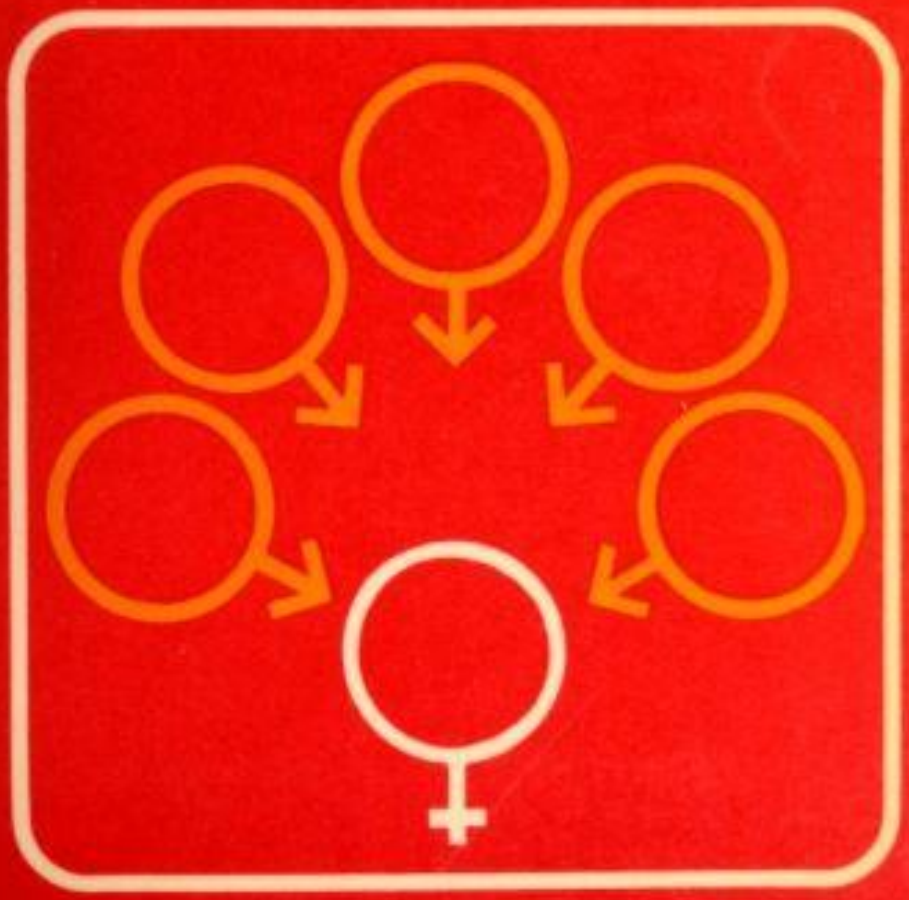
STATISTICS IN SMALL DOSES

W. M. CASTLE



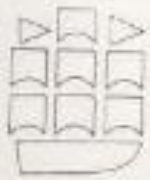
A LIVINGSTONE MEDICAL TEXT

STATISTICS IN SMALL DOSES

STATISTICS IN SMALL DOSES

Winifred M. Castle M.B., B.S., A.I.S., F.S.S.

LECTURER IN MEDICAL STATISTICS
UNIVERSITY OF RHODESIA



The Williams & Wilkins Company
Baltimore

ISBN 0 443 00734 9

Printed in Great Britain

INTRODUCTION

As a doctor I know that many of my medical and paramedical colleagues cognise in themselves a "statistical deficiency". My aim, in writing this ok, is to contribute towards a correction which I hope is both effective and latable. The examples have a medical bias but have been used and derstood by non-medical people.

I have covered the topics which appeared to me to be most useful to my lleagues. These are the description and presentation of data, ideas about iability and significance tests and the conclusions which may be drawn. e book is intended for those who want to make a start along statistical ths. This being so I make no apologies to the purists for over-simplification the numerical facts of life.

This is mainly a programmed learning text. The information is dispensed in all doses called "frames" with questions and answers to test continuously e reader's grasp of the subject matter. In order to derive maximum benefit e reader is advised to attempt to answer the questions in each frame before oking at the given response. For this purpose a cover for the answer column provided.

This book has been developed over four years and its current form is its h amended version. Criticisms and advice have been received and taken into count at every stage and I would like to thank all those who read the text conscientiously during its development. To a large extent these people ve written this book for you. I thank them, at the same time exonerating em from blame. They include medical students and colleagues at the Royal ee Hospital Medical School and at the Universities of Leeds, Pretoria and e Witwatersrand. In the University of Rhodesia, besides considerable assist·ce from within the Faculty of Medicine, groups of undergraduates and aduates in the Faculties of Science and Education have studied and com·ented upon the early versions.

I am particularly grateful for the considerable measure of assistance ovided initially by Dr David Hawkridge, now at The Open University, and bsequently by Mr Denzell Russell currently a Lecturer in the Institute of dult Education at this University. Within the Medical Faculty I owe a great al to the forbearance of the departmental secretaries, in particular Miss osemary Hore, and to the encouragement of my colleague Dr D. A. W. ttey. I am indebted to the Head of my Department, Professor W. Fraser oss, for supplying the original idea, considerable support and large slices of stationery vote throughout this venture.

1971 W. M. CASTLE

CONTENTS

EXPLANATION OF SYMBOLS

c	The number of columns in a contingency table
D	The difference between rankings
E	Expected results
f	The number of degrees of freedom
k	The number of classes (for χ^2 'goodness of fit')
μ	(mu) The population mean
N	The number of results
O	Observed results
p	Probability. In particular the probability of obtaining an equally extreme or more extreme results by chance.
r	The number of rows in a contingency table. Also Pearson's correlation coefficient.
ρ	(Rho) Spearman's correlation coefficient
Σ	(Capital sigma) Add together
σ	(Sigma) Standard deviation in the population
σ^2	Variance in the population
s	Standard deviation in a sample
s^2	Variance in a sample
$s_{\bar{X}}$	Standard error of the mean
t	Students t – for tests involving small samples and quantitative data.
χ^2	(Chi squared) For tests involving qualitative data.
X	An individual result
$\bar{X}$	The mean of the result
$(X-\bar{X})$	The deviation from the mean
Y	An individual result
z	The number of standard deviations from the mean
$>$	Bigger than

Part I Describing Numbers

Chapter 1 TYPES OF RESULTS

INTRODUCTION

This chapter teaches you to distinguish between the two types of numerical data, as this distinction is used over and over again during the book. Rates, ratios, proportions and percentages are described briefly.

1.1 The height of a patient is *measured*.
The number of patients attending a particular clinic is *counted*.
Is weight measured or counted?
What about time of survival?
Are people who have been vaccinated usually measured or counted?

Measured.
Measured.
Counted.

1.2 Results which are measured are called continuous or *quantitative*. Each individual has one measurement from a continuous spectrum, e.g. 4'11", 5'7", 6'2". Results which count people into groups with certain attributes are called discrete or *qualitative*.
Sex is .....data.

Qualitative.

1.3 Sorry to have to describe sex as 'qualitative data' but that's life!
What kind of result is heart size?

Quantitative (because it is measured).

1.4 What kind of result is red cell volume?

Quantitative.

1.5 Why is it quantitative?

Because it is measured.

1.6 Blood groups are.....data.

Qualitative, patients are counted into particular groups.

1.7 Write down 3 different sources of quantitative measurement, e.g. bladder capacity.
(1)..... (2).....
(3).....

I.Q., Examination Results and Bank Balance are 3 possibilities.

1.8 Sometimes examination results are listed qualitatively rather than quantitatively. How is this done?

Candidates are grouped on a pass/fail basis only.

1.9 Country of origin of doctors practicing in Country X.

Country	*Males*	*Females*	*Total*
England & Wales	142	28	170
Ireland	29	2	31
Italy	5	3	8
Scotland	74	15	89
S. Africa	157	26	183
U.S.A.	5	1	6
Others	7	10	17
Total	419	85	504

What type of data is this?

Qualitative.

1.10 How many doctors originated in England and Wales?

170

1.11 Qualitative results are often given as a ratio, a proportion or a percentage. Any group we single out for mention can be said to have the 'characteristic mentioned'.
If we call this group the sheep, the others are the goats.
Which are the goats in the last frame?

Doctors in X qualified outside England and Wales.

1.12 A *ratio* may be defined as

$$\frac{\text{the number of sheep}}{\text{the number of goats}}$$

In Frame 1.9 what is the ratio of

$$\frac{\text{those who originated in England Wales?}}{\text{those who did not}}$$

$$\frac{170}{334}$$

334 is the number originating outside England and Wales.

1.13 What is the ratio of women to men doctors in Country X?

$\frac{85}{419}$

1.14 If $\frac{\text{the number of sheep}}{\text{the number of goats}}$ is a ratio,

and $\frac{\text{the number of sheep}}{\text{the number of sheep and goats}}$ is a proportion,

$\frac{\text{the red cell volume}}{\text{the total blood volume}}$ is a .....

Proportion.

1.15 The ratio of women doctors to men in this country is $\frac{85}{419}$

What is the proportion?

$\frac{85}{419 + 85} = \frac{85}{504}$

1.16 $100 \times \frac{183}{504}$

is the *percentage* (%) of all doctors in Country X who originated in South Africa. The percentage is defined as 100 times the .....

Proportion.

1.17 Is the percentage therefore

$100 \times \frac{\text{sheep}}{\text{goats}}$?

No. % =

$100 \times \frac{\text{sheep}}{\text{sheep and goats}}$

1.18 Give names to the following indices from Frame 1.9 (They all refer to the U.S.A.)

(a) $\frac{1}{85}$ (b) $\frac{1}{84}$ (c) $\frac{100}{85}$

The (a) proportion (b) ratio and (c) percentage of women doctors who qualified in the U.S.A.

1.19 It is perfectly in order to cancel the numerator and denominator where possible.

$\frac{100}{85}$ % can be written .....

$1\frac{3}{17}$ % or 1.2%
This cancelling is often misused.

1.20 Ratios, proportions and percentages may seem innocent. They are sometimes misused in journals. An ear, nose and throat surgeon proudly reports that he has 100% 5-year survival rate for patients with cancer of the throat after operation.
Are you impressed?

Not until you know how many patients are involved. He may have only operated once!
Unless the actual number involved is also quoted mistrust ratios, proportions and percentages.

1.21 A normal blood picture should contain about 5 million red cells and 5 thousand white i.e. a ratio $\frac{1000}{1}$

Mr Van der Merwe has his cells in this ratio. Is his blood picture O.K.?

You can't say. He may have 5 million white cells and leukaemia. A ratio on its own is not enough!

1.22 Besides ratios, proportions and percentages, doctors quote 'rates'. For example, the Still-birth Rate is

$$\frac{\text{the number of still-births}}{\text{the total number of births}} \times 1000$$

Is this a percentage, a proportion or a ratio?

None. It is 10 times the percentage.

1.23 Some of the so-called 'rates' in Medicine are not true rates, a rate being defined as

$$\frac{\text{the number counted over a certain period}}{\text{the total at a given time}}$$

The Crude Death Rate is calculated from

$$\frac{\text{the number of deaths in a year}}{\text{the estimated population on July 1st}}$$

Is this a proper rate?

Yes. (The given definition is usually multiplied by 1000.)

1.24 $\frac{\text{The number of deaths by accident in 1066}}{\text{The number of survivors in 1066}}$ is a rate/ratio/proportion?

Ratio.

1.25 $\frac{\text{Sheep}}{\text{Total}}$ is a .....

Proportion.

$\frac{\text{Sheep}}{\text{Goats}}$ is a .....

Ratio.

$\frac{\text{Sheep counted over a period}}{\text{Total at an instant}}$ is a .....

Rate.

1.26 Ratios, rates, proportions and percentages are all means of expressing what kind of data?

Qualitative.

1.27 If you are presented with any of these values you should also be told what?

The number involved in the survey or trial.

1.28 Occasionally in medical journals the term ratio, proportion and rate are confused. One may read that the ratio of A to B is 10% instead of one to .....

Nine

$$\frac{A}{A+B} \times 100 = 10\%$$

$$\therefore \frac{A}{A+B} = \frac{1}{10}$$

$$\therefore 10A = A+B$$

$$\therefore 9A = B$$

$$\therefore \frac{A}{B} = \frac{1}{9}$$

1.29 The proportion of A is defined as $\frac{A}{?}$

Total, say A+B

1.30 If the proportion of A *increases* and this is

$$\frac{A}{A+B}$$

the proportion of B *must* increase/decrease at the same time.

Decrease

1.31 I read in a journal:
'It is particularly interesting to see that as the proportion of live eggs increases the proportion of dead eggs decreases'.
Comment.

It is inevitable rather than interesting. Like a salary cheque, as the proportion of tax increases the proportion of spending money *must* decrease. Keep your eyes open and you will notice this type of mistake in the journals.

1.32 Before learning about how to present qualitative results in journals (and how not to!) in the next chapter, check that you have learned all the main points in this chapter in the Revision Summary below.

SUMMARY

Results which are counted into groups are called qualitative.
Quantitative values are measured. The difference is important from the statistical point of view.
Qualitative data can be summarised as

ratios $= \frac{\text{sheep}}{\text{goats}}$ and proportions $= \frac{\text{sheep}}{\text{total}}$ where the percentage (%) is

100 times the proportion.
A rate is similar to a proportion but its denominator is a static measureme
whereas the numerator is counted over a period of time.
It is wrong to state one of these indices without quoting the number involv
It is wrong to take much notice of them in the absence of this information
Keep a look out for their misuse.

Chapter 2 ILLUSTRATING COUNTS

INTRODUCTION

Sometimes research workers spend a lot of their time obtaining results and then present them poorly in journals. Illustrating data is a useful topic as badly presented results can mislead you. Diagrams should aid the reader by saving him time and by highlighting the points. Qualitative and Quantitative data are presented differently.

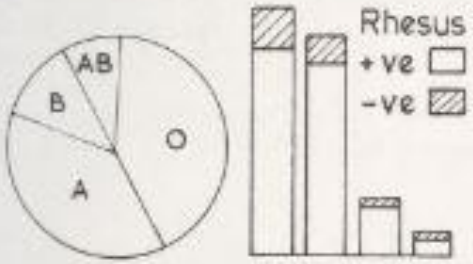
2.1 There are 4 ways of presenting qualitative data.

(a) Pie diagram
(b) Pictogram
(c) Bar chart
(d) Proportional bar chart

These 4 methods are used below to illustrate the fact that

45% of Europeans have blood group O
41% have blood group A
10% have blood group B
4% have blood group AB

Can you label the methods? (You will need to rely on an inspired guess).

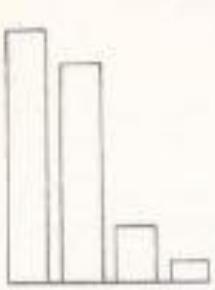


a b

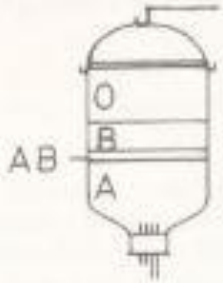
(a) is a ..... (b) is a .....

contd. overleaf

2.1 *contd.*



c



d

(c) is a ..... (d) is a .....

(a) Pie diagram
(b) Proportional bar chart
(c) Bar chart
(d) Pictogram

If your answers are correct go straight to Frame 2.5.

2.2 Why is (a) called a pie diagram?

It is pie-shaped.

2.3 Why is (c) called a bar chart?

It is shaped like a series of bars.

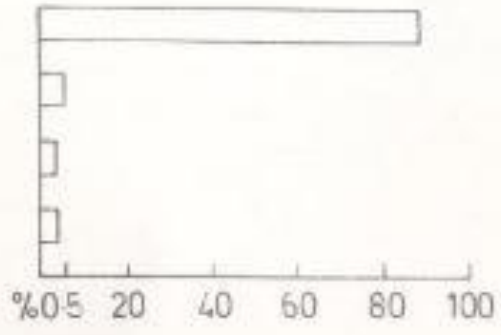
2.4 Why is (b) called a proportional bar chart?

It is a bar chart with sub-divisions proportioned off.

2.5 (d) is a pictogram. The 'gram' infers measurement and 'picto' a picture. What picture would you use to represent the composition of cow's milk by volume?

I asked you what you would use - you are entitled to choose. A milk bottle or urn are possibilities. Some Picasso amongst you may choose to draw a cow.

2.6 What method of representation is this?



Bar chart - horizontal this time.

2.7 Frame 2.6 represents some results. Is there any way of knowing what they are about?

There is no way of knowing unless you are psychic.

2.8 One principle in illustrating results is to give the diagram a .....

Heading or title.

2.9 The diagram in Frame 2.6 depicts the composition of cows' milk by weight.

88% is water
3% is protein
4% is fat
5% is carbohydrate

Besides labelling the heading what else should be labelled?

The various sections.

2.10 The diagrams below give 2 further labels besides the heading and components. What are they?

The author and the date.

Areas in which doctors in Country X are practising. Results collected by Professor W. F. Ross, 1967.

2.11 One of the principles of presenting data is to label fully. What should be labelled?

(1) (2) (3) (4)

Heading, components, author, date.

2.12 To recap, what are the 4 main ways of presenting diagrammatically data which is qualitative?

Pie diagram, pictogram, bar chart and proportional bar chart.

2.13 Besides being *fully labelled*, or self explanatory, diagrams must be easy to understand. They should be as simple as possible, and not over-crowded with information. Look at the .... below.

How would you improve it?

Areas in which doctors in Country X are practising (or are retired) showing sex. Data collected by Professor W. F. Ross, 1967 (but not presented like this!)

Pie diagram.

There is too much on one diagram. Split it into two.

2.14 *Simplicity* is the second principle in presenting results. A good diagram will save a lot of words in the text. The diagram in Frame 2.10 is/is not a simple diagram because it would/would not save a lot of words in the text.

Is.
Would.

2.15 What may be described as the first 2 principles in depicting data?

Full labelling.
Simplicity.

2.16 There is a third - *honesty*. I don't mean just telling the truth but also the whole truth and nothing but the truth. There must be no attempt to mislead and justice must be seen to be done. A manufacturer of a hair shampoo, 'Coo,' gave free samples to 10 filmstars. 1 lost the sample and the other 9 used theirs. The manufacturers claim *'9 out of 10 filmstars use 'Coo' shampoo.'* Is he right?

You say yes.
The slogan is dishonest and misleading although it contains the truth, I suppose.

You say no.
You are not gullible you are probably a very good cheat.

2.17	How can you be dishonest with proportions and ratios, etc.?	By failing to record the total number considered.
2.18	The same principles of presentation apply to quantitative data and moreover it is easier to be dishonest with that type of data. Apart from honesty, what other principles are involved in illustrating data?	Full labelling and simplicity. Full labelling and simplicity.
2.19	One of the reasons for diagrammatic presentation of data is to make the points clearer. What is the other?	To save words.
2.20	What methods do you know for illustrating counts?	Pie diagram. Bar chart. Proportional bar chart. Pictogram.
2.21	Why would you use these methods?	To save words and make the points clear.
2.22	To what principles would you adhere?	Full labelling, simplicity and honesty.

2.23 **Practical Example**

Use a proportional bar chart to present the data in Frame 1.9

SUMMARY

I hope your example is fully labelled and is as simple as possible. Honesty is the other criterion but this will be followed up further in the next chapter.

To present qualitative data diagrammatically, pie diagrams, pictograms, bar charts and proportional bar charts are used. They are intended to save words and make points clearer.

Chapter 3 ILLUSTRATING MEASUREMENTS

INTRODUCTION

Quantitative results are generally more cumbersome than qualitative to represent diagrammatically. It is easier to represent sheep and goats than to distinguish diagrammatically a 200 lb. sheep from a 240 lb. sheep.

3.1 *Birth weights of 12 babies of mothers found to be suffering from sugar diabetes*

(Fictitious data)

103 oz.	131 oz.	143 oz.
114 oz.	138 oz.	146 oz.
114 oz.	138 oz.	151 oz.
122 oz.	138 oz.	170 oz.

What kind of data is this?

Quantitative.

3.2 The characteristic which is varying is called, not surprisingly, the *variable*. What is the variable in the last frame?

The birth weight of babies of diabetic mothers.

3.3 Such results as in Frame 3.1, from large samples, are grouped before illustration. See below. Is there more information in the data after grouping?

No. For example no distinction is made now between the 122 oz. and 138 oz. birth weights.

Data from Frame 3.1 Grouped.

Group No.	*Group*	*Frequency*
1	80 – 99 oz.	0
2	100 – 119 oz.	3
3	120 – 139 oz.	5
4	140 – 159 oz.	3
5	160 – 179 oz.	1
6	180 – 199 oz.	0

3.4 The price to pay for being able to illustrate measurements is loss of some information. The number in each group is called the *frequency* of that group. What is the frequency in the group 120 – 139 oz. in the last frame?

5

3.5 All the frequencies considered together form the *frequency distribution*. The frequency distribution in Frame 3.3 is 0, 3, 5, 3, 1, 0. What is the frequency distribution below?

Birth weights of 16 babies of normal mothers (Fictitious data).

52 oz.	103 oz.	109 oz.	127 oz.
79 oz.	104 oz.	111 oz.	149 oz.
80 oz.	104 oz.	120 oz.	150 oz.
100 oz.	106 oz.	121 oz.	162 oz.

The above data Grouped.

Group No.	*Group*	*Frequency*
1	20 – 39 oz.	0
2	40 – 59 oz.	1
3	60 – 79 oz.	1
4	80 – 99 oz.	1
5	100 – 119 oz.	7
6	120 – 139 oz.	3
7	140 – 159 oz.	2
8	160 – 179 oz.	1
9	180 – 199 oz.	0

0, 1, 1, 1, 7, 3, 2, 1, 0

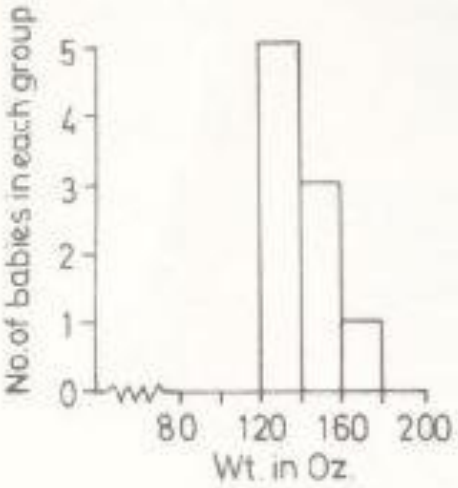
3.6 What is the variable in the last frame?

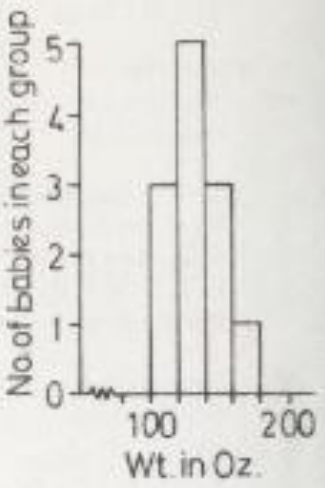
Birth weight of babies of normal mothers.

3.7 Why is haemoglobin level a variable?

Because it varies.

3.8 Having grouped the data we can present it more easily in a diagram. Below the results from Frame 3.3 are partly filled in. Complete the diagram.





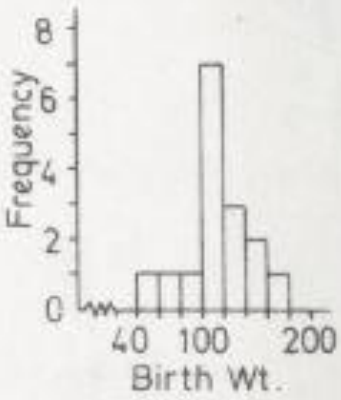
3.9 This method of illustration is called a *histogram*. Instead of 'No. of babies in each group' we could write .....

Frequency.

3.10 In your own words describe a histogram, and what it is used for.

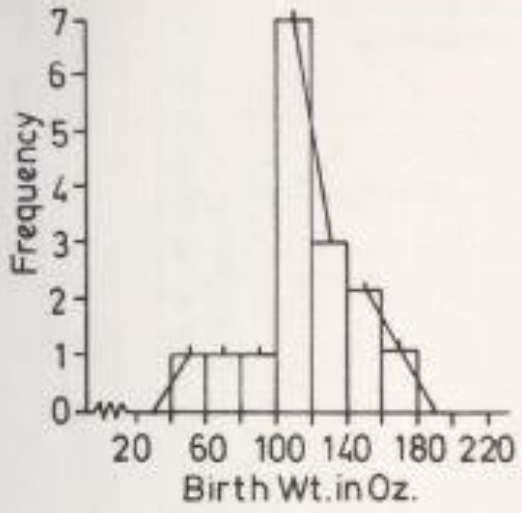
A histogram is a method of presenting quantitative data. Along the horizontal axis is the variable and up the vertical axis is the frequency. The histogram is a series of boxes standing side by side. The size of the box indicates the frequency in the group.

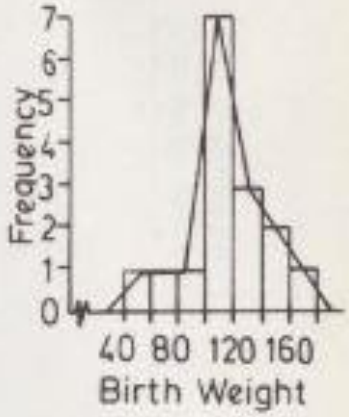
3.11 Draw a histogram to illustrate the data in Frame 3.5.



3.12 Another method of presenting measured data is the *frequency polygon*. If the mid-points of the box lids are joined by a series of straight lines in the histogram you have just drawn, you have the equivalent frequency polygon.

Complete:-





3.13 The boxes are not shown on a frequency polygon only the series of straight lines. If you wanted to show two frequency distributions such as Frames 3.3 and 3.5 on the same diagram, would you use 2 histograms or 2 frequency polygons?

2 frequency polygons, otherwise the boxes would overlap.

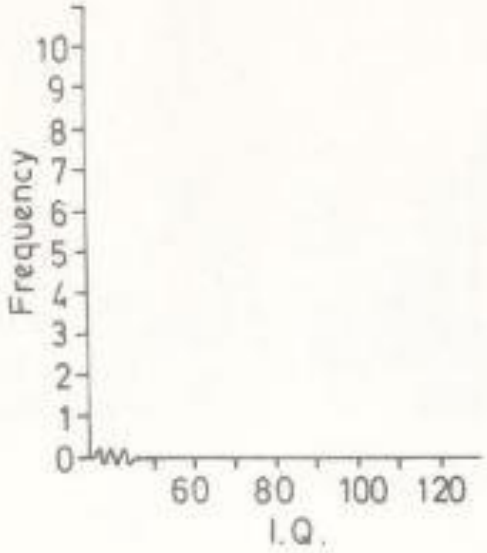
3.14 The frequency polygon is always continued until it meets the horizontal axis. Put another way, the frequency of the outside groups included in a frequency polygon is always .....

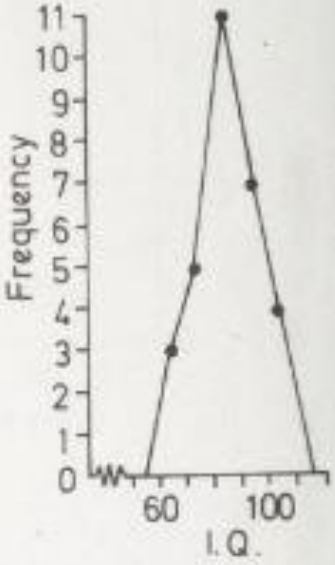
Zero.

3.15 30 spastic children had the following I.Q.'s:

I.Q.	*Frequency*
60-69	3
70-79	5
80-89	11
90-99	7
100-109	4

Construct below a frequency polygon to represent this frequency distribution.





3.16 What are the 3 principles for good presentation of all results?

Full labelling; simplicity; honesty.

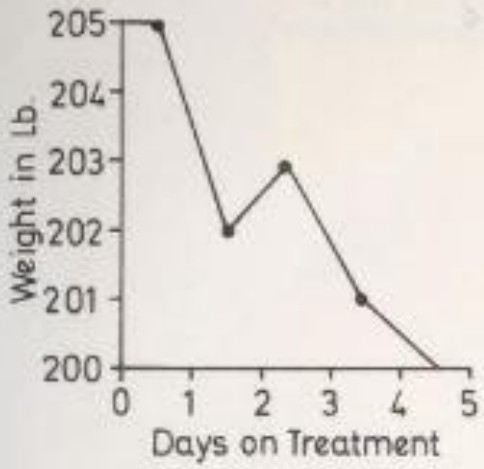
3.17 What is the commonest dishonest method used with qualitative data?

Not stating how many results were included in quoting percentages etc.

3.18 With quantitative data, cheating is very easy. I will teach you 3 tricks. 1 such trick is to *suppress the zero*.

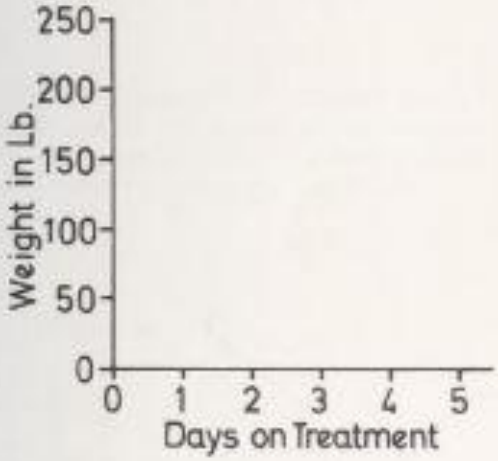
e.g. *Weight loss on Silfy tablets.*

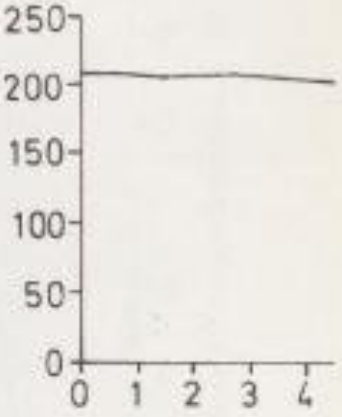
3.18 *contd. from opposite page*



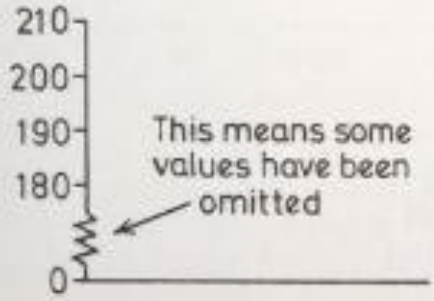
Re-sketch the diagram without suppressing the zero (i.e. mark 0 on both axes).

Not so impressive is it?





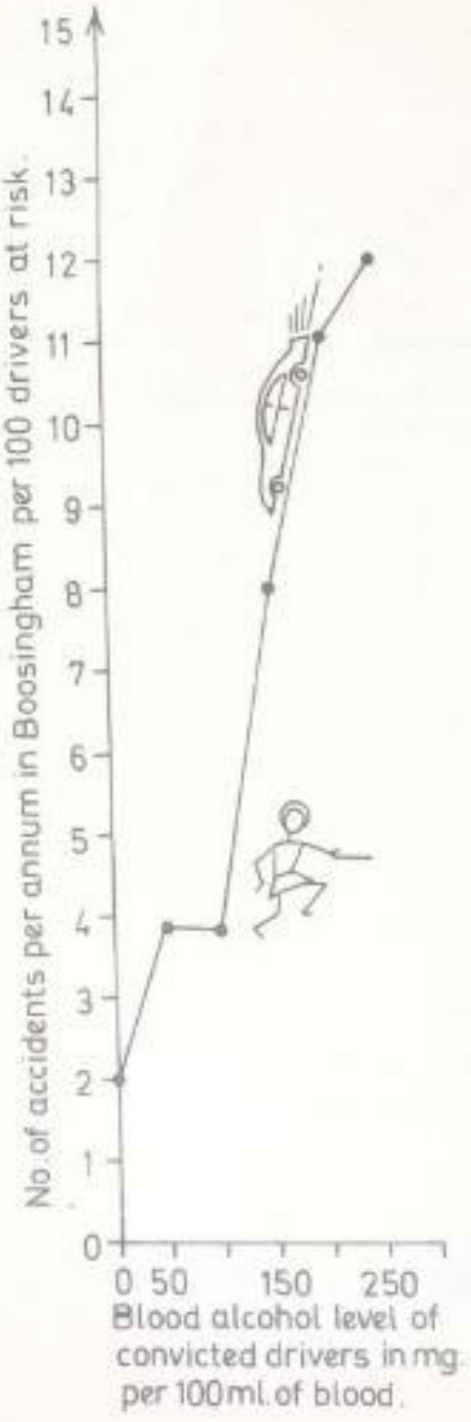
3.19 Although zero must always be shown on a diagram, sometimes the axis can be condensed as follow:-



contd. overleaf

3.19 *contd.*

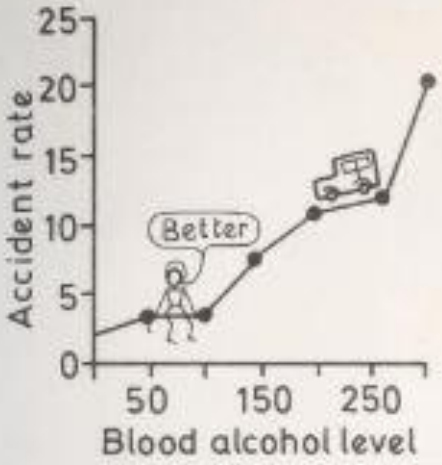
Closely related to suppressing the zero is the trick of *inflating or exaggerating the scale*, e.g.



contd. on opposite page

3.19 *contd.*

On a more reasonable scale this becomes:

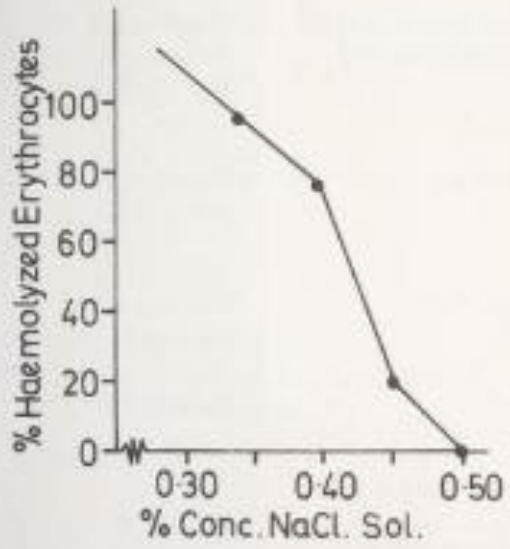


What is a reasonable scale?

A reasonable scale is one which neither over-emphasises nor under-emphasises the evidence.

3.20 What is wrong with the diagram below? It shows the percentage of erythrocytes haemolysed in various concentrations of salt solution (Wintrobe's method).

It infers that more than 100% of the red cells survived at low concentrations. This is impossible.

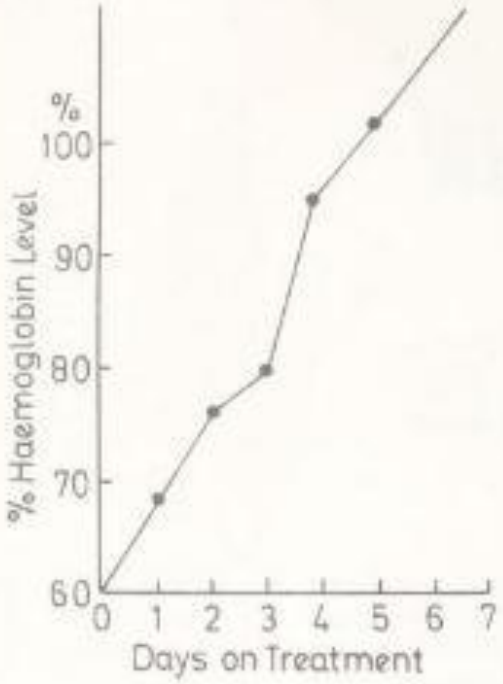


3.21 The trick in Frame 3.20 is technically called *extrapolation*. (extending the line beyond the actual results.)
What other 2 tricks do you know?

Suppressing the zero.
Inflating the scale.

3.22 What tricks have been employed in this diagram showing a patient's increase in haemoglobin level after therapy with the drug 'Ironical.'

All 3.



3.23 Which are the 3 commonest tricks used to illustrate quantitative data dishonestly?

Suppressing the zero.
Inflating the scale.
Extrapolation.

3.24 What methods do you know for illustrating measurements?

Histograms and Frequency polygons.

3.25 What methods do you know for illustrating counts?

Pie diagrams.
Bar charts.
Proportional bar charts.
Pictograms.

3.26 The difference between a bar chart and a histogram is that the ..... is used when data is measured and the ..... for ..... data. In the histogram the groups adjoin each other and the boxes approximate each other, whereas with the bar chart the groups are usually .....

Histogram.

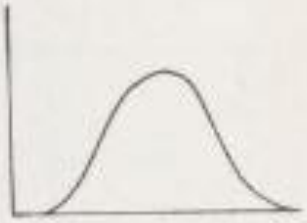
Bar chart, Qualitative.

Separate.

3.27 Ideally between 10 and 25 classes should be used in a histogram. How many are used here?

18

3.28 Sketch the frequency polygon that would be constructed from the histogram in the preceding frame - it would look like a curve.



This diagram is the subject matter of the next chapter.

3.29 **Practical Example**

Sketch the data from Frames 3.3 and 3.5 on the same diagram.

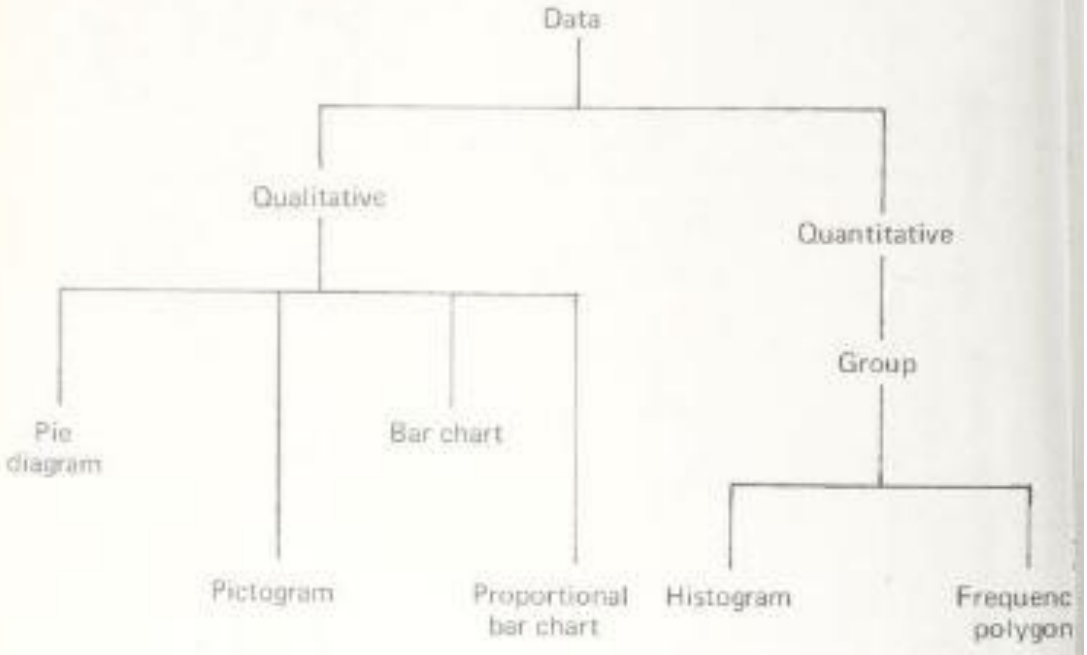
Comment.

Babies of diabetic mothers seem to be bigger.

contd. overleaf

SUMMARY

Before presenting measurements the results are grouped. The number in each class is called the frequency. The frequency in all the classes is called the frequency distribution. This grouping makes illustration easier, but the price in some loss of detail. Ideally, 10 to 25 different groups can be used.
The frequency distribution is represented by a histogram or frequency polygon. When 2 or more frequency distributions are superimposed it is better to use 2 frequency polygons.
The principles for presenting all data are full labelling, simplicity and honesty. The commonest tricks with quantitative data are suppression of the zero, inflating the scale and extrapolation.



Chapter 4 THE NORMAL DISTRIBUTION

INTRODUCTION

In the last section we imagined the shape of a curve constructed from a particular histogram. Most biological and medical sources of quantitative results either follow that curve or can be modified easily so that they do. Because it occurs so commonly it is important to understand it. (Without its formula, that is!)

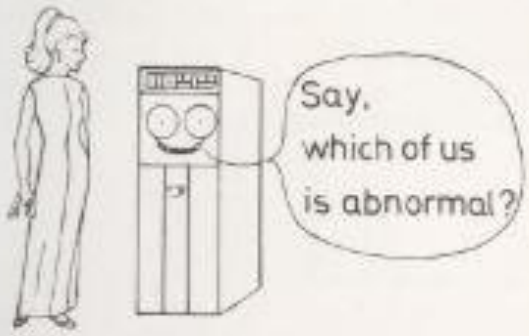
4.1 A factor which varies is called a ..... Variable.

4.2 In the pure sciences such as Physics and Chemistry there is not so much inherent variability as in Biology and Medicine. One chemical carbon atom is much like another carbon atom - but when they are biologically arranged the effects can range from 'stunning' to 'mediocre' and even beyond.
How tall are you? Your height is .....

4.3 I am 5 ft 3 in. Are you the same height as I am? Probably not.

4.4 If your height is different from mine, which of us is abnormal? Neither of us - so far as height is concerned!

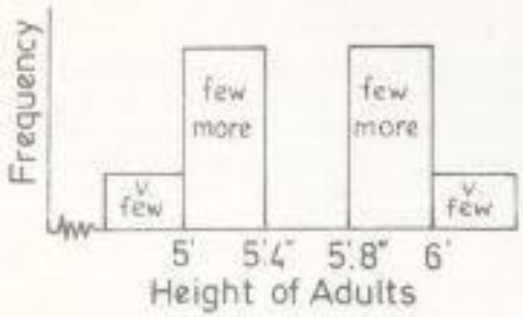
4.5



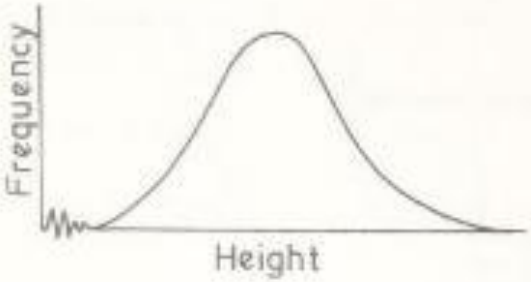
4.6 Do you know more people over 6 ft tall than under 6 ft tall? No.

4.7 Do you know more adults less than 5 ft tall than over 5 ft tall? No.

4.8 The majority of adults are between 5 ft 4 in and 5 ft 8 in tall. Complete the histogram.



4.9. If we take many more groups and sketch the curve for people's height, it could look like this:

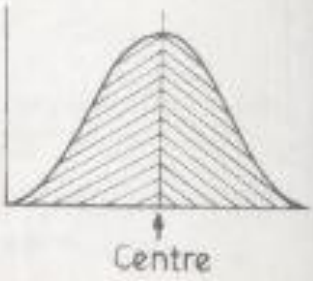


This is called the *normal distribution*.
It is shaped like a .....
It applies to qualitative/quantitative data.

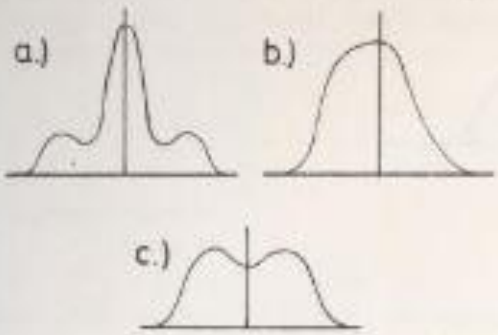
It is often described in textbooks as being bell-shaped.
Quantitative.

4.10 A symmetrical curve is one with 2 sides of the centre absolutely corresponding. The curve of the normal distribution is/is not symmetrical.

Is.



4.11 Which of the following are symmetrical?

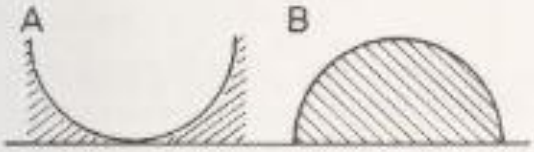


(a) and (c) are.

4.12 Describe the normal curve.

It is bell-shaped and symmetrical.

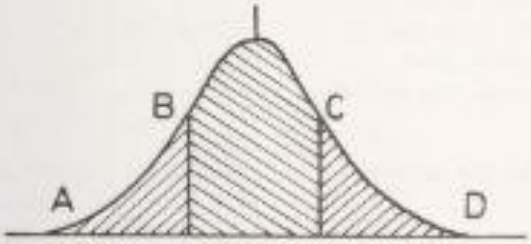
4.13 Relative to the base-line a convex surface is arched and a concave surface is hollow.



A is ..... B is .....

Concave. Convex.

4.14 In this diagram indicate the part which is convex and the parts which are concave.

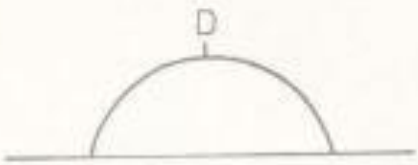


Between B and C it is arched or convex, otherwise it is hollow or concave.

4.15 The point where a convex section of a curve changes to a concave section is called a *point of inflection*.
How many points of inflection has a normal distribution curve?

2 B and C in the last frame.

4.16 D is/is not a point of inflection?

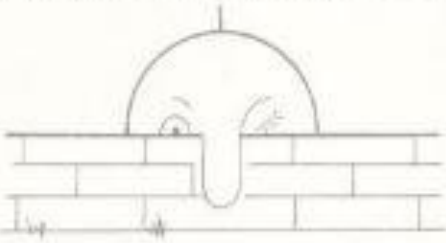


Is not.

4.17 Why?

The curve remains convex at point D.

4.18 How many points of inflection has Chad?



2 only - on his nose.

4.19 This centipede has ..... points of inflection and is/is not symmetrical.

4 on each surface.
Is not.

4.20 For a variable to have a frequency distribution like the normal distribution the majority/minority have a measure near the middle and the majority/minority have measures near the extreme.

Majority.

Minority.

4.21 Is income distributed like the normal distribution?

No. The majority have a small wage and the minority (like doctors) a large wage.

4.22 However, many variables based on living objects can be approximately described as being normally distributed. Height is one example, write down 3 others.

1. .....
2. .....
3. .....

I.Q., Weight, Bladder Capacity, Haemoglobin level, Examination marks usually, etc.

It is fairly hard to think of variables which are not approximately normally distributed.

4.23 Why do you think the *normal* distribution is so called?

Because it is the distribution normally encountered.

4.24 Describe the normal curve.

Symmetrical; bell-shaped; 2 points of inflection.

SUMMARY

Variation between individuals is a natural phenomenon, thank goodness. The fact that most variables either follow the normal distribution directly or can easily be adjusted to do so is very useful and enables numbers to be used to answer theories. The normal distribution is symmetrical and bell-shaped and has 2 points of inflection where the shape of the curve changes from convex to concave or vice versa.

Chapter 5 NOTATION

INTRODUCTION

You are now to be introduced to symbols which will save a lot of words later. If your arithmetic is such that you understand this chapter's summary you can skip the chapter.

5.1 N is used for the number of results.
What is N in Frame 3.1?

12

5.2 Usually X is used instead of an individual result. What is the last X in Frame 3.1?

170 oz.

5.3 If we have two results for each patient e.g. a height and a weight, we usually call one X and one Y.
X and Y then represent two factors which vary or are .....

Variables.

5.4 How many X's are there in any set of results?

N

5.5 Σ is capital 'S' in Greek. It is pronounced 'sigma'
Σ means *add together* all the results.
What does Σ (1,2,2,3) equal?

8

5.6 ΣX means add together all the values of X.
It is pronounced .....X.
What is ΣX for these 5 fictitious haemoglobin levels?

Sigma

80, 90, 100, 110, 120.

500

5.7 Below are corresponding values of X and Y for 5 students. X is haemoglobin level and Y is intelligence quotient, if you like.
N in each case equals .....
ΣY equals .....

5
550

X 80, 90, 100, 110, 120
Y 80, 90, 100, 110, 170

5.8 What do $\frac{\Sigma X}{N}$ and $\frac{\Sigma Y}{N}$ equal in the previous frame?

$\frac{\Sigma X}{N} = \frac{500}{5} = 100$

$\frac{\Sigma Y}{N} = \frac{550}{5} = 110$

5.9 $\frac{\Sigma X}{N}$ is the average of the X variable

Its symbol is $\bar{X}$, called X bar.

What is the result $\frac{\Sigma Y}{N}$ called?

The average of Y or $\bar{Y}$ or Y bar.

5.10 The average in statistical jargon is usually called the *mean*.

$\bar{Y}$ is the ..... of all the Y.

Mean.

5.11 $\bar{X}$ in Frame 5.6 = 100

What does $\sqrt{(\bar{X})}$ equal?

10

5.12 If 4 values of X are 2,4,6,8, which symbol equals 4?

Which symbol equals 5?

Which symbol equals 20?

N = 4

$\bar{X}$ = 5

ΣX = 20

5.13 When you see a capital sigma you do what?

Add the results together.

5.14 Sometimes we will want the results squared before adding.

We then write $\Sigma(X^2)$ or $\Sigma(Y^2)$

If Y is 1,2,2,3,

$\Sigma(Y^2)$ = ...

1+4+4+9 = 18

You perform the task in brackets first; that is what the brackets mean.

5.15 $(X - \bar{X})$ means what?

An individual result minus the mean.

5.16 $(X - \bar{X})$ is called *the deviation from the mean*. What is the last value of the deviation from the mean in Frame 5.6?

$120 - 100 = +20$

5.17 It is a rule that you perform the task shown in brackets first.
What is $\Sigma(X - \bar{X})$ in Frame 5.6?

$(-20)+(-10)+(0)+(+10)+(+20) = 0$

5.18 $\bar{Y}$ in Frame 5.7 is 110.
What is $\Sigma(Y - \bar{Y})$ in that frame?

0

5.19 $\Sigma(X - \bar{X})$ and, of course, $\Sigma(Y - \bar{Y})$ always equals zero.
In words, ...
.....equals zero.

The sum of the deviations from the mean.

5.20 In algebra XY equals a value of X multiplied by its corresponding value Y. If X is 1,2,3 while the 3 corresponding values for Y are 1,3,4 the 3 values of XY are .....
and $\Sigma(XY) =$ .....

$1 \times 1 = 1 \quad 2 \times 3 = 6,$
$3 \times 4 = 12$
$\Sigma(XY) = 1+6+12$
$= 19$

5.21 Remember, always perform the task in brackets first.
If X is 1,2,2,3,
$\bar{X}$ equals .....
and $\Sigma(X - \bar{X})^2$ equals .....

2
$(-1)^2+(0)^2+(0)^2+(+1)^2 = 2$
Remember a negative number squared gives a positive answer.

5.22 If Y is 0,2,3,3
$\bar{Y}$ equals .....
and $\Sigma(Y - \bar{Y})^2$ equals .....

2
6

5.23 Meanwhile $\Sigma(Y - \bar{Y})$ = ...

0 as always.

5.24 If Y is 0,2,3,3

ΣY = ...

8

$\Sigma(Y^2)$ = ...

22

5.25 Do you think $(\Sigma Y)^2$ in the last frame equals (A) 22
or (B) 64
or (C) any other answer?

You say A - go to Frame 5.26.
You say B - go to Frame 5.27.
You say C - go to Frame 5.26.

5.26 You are incorrect.
$\Sigma(Y^2)$ is not the same as $(\Sigma Y)^2$

In $\Sigma(Y^2)$ you square and then add, while in $(\Sigma Y)^2$ you add the results and then square your answer; (remember, perform the bracket first)

If Y is 0,1,2

$\Sigma(Y^2)$ equals .....+.....+..... which equals 5.

$0 + 1 + 4$

$(\Sigma Y)^2$ equals $(.....+.....+.....)^2$ which equals 9.

$(0 + 1 + 2)^2$

5.27 You know the difference between $\Sigma(Y^2)$ and $(\Sigma Y)^2$

If X is 0,1,2 and Y is 1,2,3

$(\Sigma X)(\Sigma Y)$ = ..., and

$(\Sigma X)(\Sigma Y) = 3 \times 6 = 18$

$\Sigma(XY)$ = ...

$\Sigma(XY) = 0 + 2 + 6 = 8$

5.28 This is all the notation you need to know throughout this book.

To revise it,

where X equals 1, 2, 4, 5
and Y equals 0, 0, 2, 2

N in both cases	equals	4.
ΣX	equals	12.
ΣY	equals	4.
$\bar{X}$	equals	3.
$\bar{Y}$	equals	1.
$\Sigma(X^2)$	equals	46.
$(\Sigma X)^2$	equals	144.
$\Sigma(XY)$	equals	18.
$(\Sigma X)\ (\Sigma Y)$	equals	48.
$\Sigma(X - \bar{X})$	equals	0 of course.
$\Sigma(Y - \bar{Y})^2$	equals	4.

5.29 **Practical Example**

Choose any 10 different values for X for the table below (i.e. N = 10). Using these numbers calculate in the spaces provided:

(a) $\dfrac{\Sigma(X - \bar{X})^2}{N - 1}$ =

(b) $\dfrac{\Sigma(X^2) - \dfrac{(\Sigma X)^2}{N}}{N - 1}$ =

The two answers should be equal.

contd. on opposite page

5.29 *contd.*

X	$(X - \bar{X})$	$(X - \bar{X})^2$	X	X^2
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
$\Sigma(X)$ =	$\Sigma(X-\bar{X})=0$	$\Sigma(X-\bar{X})^2=$	$\Sigma(X)$ =	$\Sigma(X^2)$ =
$\bar{X}$ =			again	

Therefore $\frac{\Sigma(X-\bar{X})^2}{N-1}$ =

and $\frac{\Sigma(X^2)-\frac{(\Sigma X)^2}{N}}{N-1}$ =

SUMMARY

N is the number of results recorded.

X or Y are individual observations.

Σ (capital sigma) is the summation sign and means 'add together.'

$\bar{X} = \frac{\Sigma X}{N}$ is the mean

$(X - \bar{X})$ is the deviation from the mean.

$\Sigma(X - \bar{X})$ always equals 0.

Always work out the results in brackets first.

e.g. $\Sigma(X^2)$ is the square of the individual results, then added together.

However, look out for the bracket including the summation sign,

e.g. $(\Sigma X)^2$ is the sum of the individual results squared.

Well done. These points will keep coming out throughout this book – rather like a rash, I suppose. Honestly, this is all the arithmetic you need to understand.

Chapter 6 MEASURING THE MIDDLE

INTRODUCTION

Rather than talk about a set of results in terms of all the individual values we can summarise the information. To do this we need to be able to describe 3 things.

1. the *shape* of the results (e.g. the normal distribution considered in Chapter 4)
2. the *middle* of this distribution (considered in this Chapter) and,
3. the *degree of variation* (considered in the next Chapter)

6.1 Describe the normal distribution.

It is symmetrical and bell shaped with 2 points of inflection.

6.2 The middle of the normal distribution is the *mean*. What is the symbol for the mean?

$\bar{X}$

6.3 The mean is the measure of the middle of the distribution as all the results tend to lie about the mean. Is the 'range' of results a measure of the centre?

No.

6.4 We are going to discuss 3 measures indicating the centre. One is the *mode*, one is the *median*, and the other is the .....

Mean.

6.5 If results are listed in order of size they are called an *array*. An examination list in alphabetical order is/is not an array.

Is not – unless the Aarons are at the top of the class and the Zvakanakas at the bottom, etc. !

6.6 Are the results given in Frame 5.6 an array?

Yes.

6.7 Do the results need to be arrayed before the mean is calculated?

No.

6.8 The median is *the middle value in an array*.
What is the median in Frame 5.6

100

6.9 The results quoted in Frame 5.7 are repeated here

X 80, 90, 100, 110, 120,
Y 80, 90, 100, 110, 170.

What are the 2 values of the median?

100 – the same in each.

6.10 The extreme value, 170 does not affect the median. Are the means in both distributions the same?

No.
$\bar{X}$ is 100
$\bar{Y}$ is 110

6.11 Not only do the extreme values affect the mean but an extreme value has a greater/lesser effect than one near to the middle.

Greater.

6.12 When we are interested in whether cases fall in the upper or lower half of the distribution, and not particularly in how far they are from the central point, we use the median/the mean.

The median.

6.13 The mean uses all the information available. Does the median?

No. It is not such a reliable measure. The result 170 does not affect it.

6.14 Which is easier to calculate, the median or the mean?

The median, especially if the results are already arrayed.

6.15 Choosing between the median and the mean, where applicable, we know:–

The ..... is more reliable and is, in fact, usually used.

Mean.

The ..... is easier to calculate than the .....

Median.
Mean.

The ..... is not affected by extreme values.

Median.

To calculate the ..... you need to array the results first.

Median.

6.16 Are these sedimentation rates an array?

7 13, 11, 15, 10, 20

No.

6.17 What is the value of N in the previous frame?

6

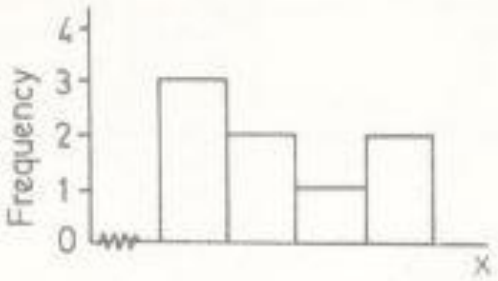
6.18 So far, when we have calculated the median, N has been an odd number so that there has only been one middle value.
Is N odd or even in Frame 6.16 ?

Even.

6.19 When N is even the median is the average or mean of the middle two values.
What is the median in Frame 6.16?
(Hint: array the results first.)

12. The average of 11 and 13

6.20



The distribution is 1,1,1,2,2,3,4,4

X is (1,1,1,2,2,3,4,4) above.
What is the mean value?
What is the value of the median?

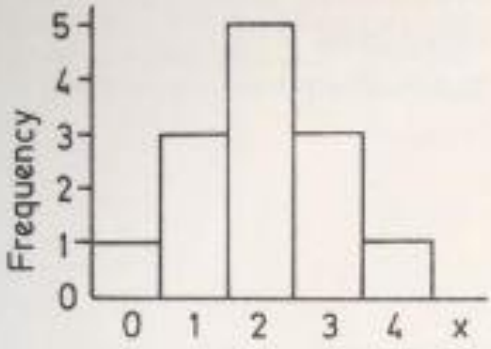
Mean = $\frac{18}{8}$ = 2¼

Median = 2

6.21 The mode is the value which occurs most frequently — the most fashionable number. What is the mode in the last frame?

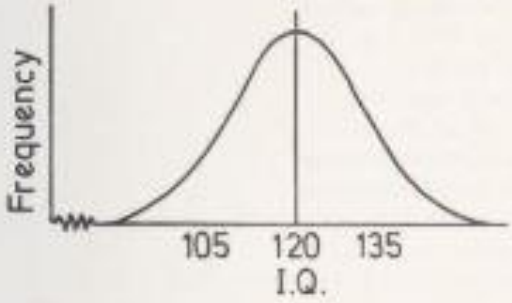
1

6.22 In this diagram the mode, median and mean are/are not the same.



Are.
The distribution is 0,1,1,1,2,2,2,2,2,3,3,3,4.

6.23 What is the value of the median, the mode and the mean in this normal distribution?



120.
All these measures of the middle are equal in the normal distribution.

6.24 The normal distribution is bell-shaped and symmetrical about the ....., ..., and .....

Mean, median, mode.

6.25 The ..... and ..... are easier to calculate than the ..... if the results are arrayed.

Mode and Median.
Mean.

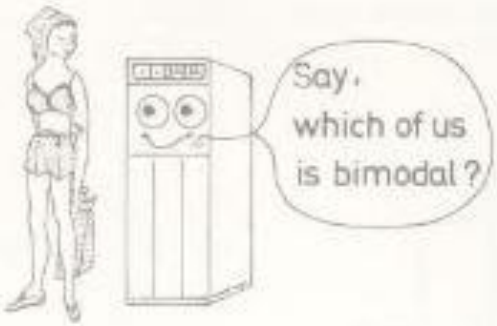
6.26 What is the mode?

The value occurring most frequently - the most fashionable - the peak on a graph.

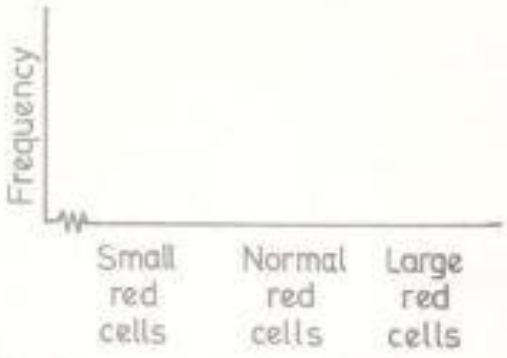
6.27 The camel can be said to be bimodal!! Why?

When it has 2 modes or humps.

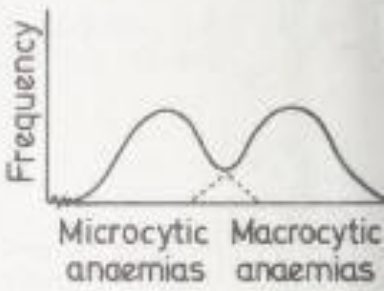
6.28



6.29 The mode is the least valuable measure of the middle. Most sets of results follow the normal distribution and the mean is the most accurate measure. However, sometimes in research one comes upon a distribution with 2 modes. This is a useful sign that 2 groups of very dissimilar people are mixed together, (statistically speaking - not an orgy!) — that the group is probably heterogeneous. For example, 2 chief groups of anaemias are macrocytic (with big cells) and microcytic (with small cells).
Sketch cell size in anaemias.



Distribution of cell size in all cases of anaemia.



Really 2 normal distributions with some overlap.

6.30 The most reliable value of the centre is the ..... It is/is not the only measure of the middle which uses all the information.

Mean.
Is.

6.31 Unless we wish to know the most typical value, in which case we would use the .....; or we wish to know whether cases fall in the top half or the lower half when the ..... is used; it is best to calculate the .....

Mode.

Median.

Mean.

6.32 In this array:

1,2,3,3,4,5,10,12,

3 is the value of the .....

5 is the value of the .....

and ..... is the value of the .....

Mode.

Mean.

3½

Median.

SUMMARY

The 3 most useful measures of the centre are the mean, the median and the mode.

The mean, or average, has the symbol $\bar{X}$ and is the most reliable measure. It is markedly affected by extreme values.

The median and mode are calculated after the results are placed in order of size in an array.

The median is the middle value in an array (or the average of the middle 2 values if N is even.) It is usually used if we are particularly interested in whether cases fall in the upper or the lower half of the distribution.

The mode is easy to distinguish, being the value which occurs most frequently. If a frequency distribution is bimodal it usually means that the group is not homogeneous and 2 very different groups are mixed together.

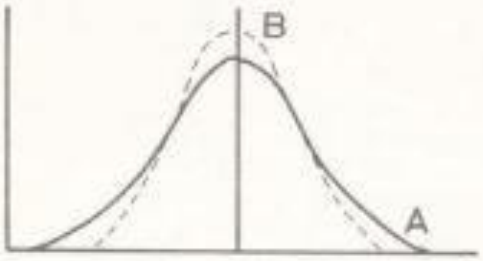
In the normal distribution the mean, the median and the mode fall in the same place.

Chapter 7 MEASURING THE VARIATION

INTRODUCTION

So far we have summarised a set of results by describing the shape and centre. In this chapter we discuss how to measure variation.

7.1 Below we have 2 normal distributions, each with the same mean. 1 set of results varies more than the other – it has a bigger scatter. Is it distribution A or distribution B?



Distribution A. The results in distribution B cluster more closely about the mean. A has more variation between results.

7.2 There were 3 measures of the centre discussed. What were they?

The mean, the mode and the median.

7.3 Similarly there are 3 measures of variation. The first is the *range*. The range is the difference between the largest and smallest result. Does it use all the available information?

No

7.4 What is the range in Frame 5.6?

120 – 80 = 40

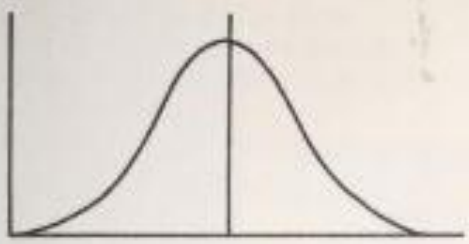
7.5 Complete the gaps.
1, 2, 3, 3, 5, 9 is a distribution where 3 is the ..... and the .....
and where 8 is the .....

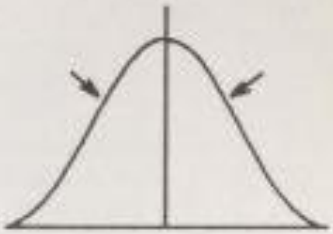
Median Mode
Range.

7.6 The mean is reliable because it uses all the information. Is the range a reliable measure of variation?

No. It does not use all the information. It is normally only used when the median is used to measure the centre.

7.7



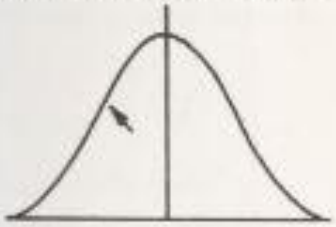


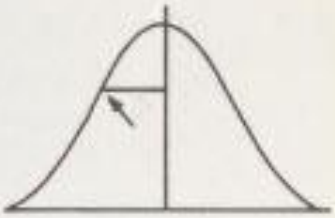
Mark the points of inflection above.

7.8 In the previous frame draw a horizontal line joining the points of inflection. This length can be used as a measure of variation. If the results are very scattered the line is longer/shorter than if they cluster around the mean.

Longer.

7.9 In fact the length of this horizontal line from a point of inflection to the *mean* is called the *standard deviation*. It is a very useful measure of variation. Draw in the standard deviation below.





7.10 The standard deviation is often given the symbol s. We need to calculate this value. We know it does not equal $\Sigma(X - \bar{X})$. Why?

This value always equals 0.

7.11 Therefore, $\frac{\Sigma(X - \bar{X})}{N}$, is also equal to 0, so cannot be used as a measure of variation either. In fact, s has a nastier formula.

contd. overleaf

7.11 *contd.*

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N - 1}}$$

s^2 is the other measure of variation and is called the *variance*.
Give the formula for the variance
s^2 =

$$\frac{\Sigma(X - \bar{X})^2}{N - 1}$$

7.12 $s^2 = \dfrac{\Sigma(X - \bar{X}^2)}{N - 1}$

You have met the $\Sigma(X - \bar{X})^2$ part before at the end of Chapter 5 and you know that N is what?

The number of results.

7.13 Where X is 0, 1, 2, 2, 5

$\bar{X}$ = ...

$\Sigma(X - \bar{X})^2$ = ...

$$\bar{X} = \frac{10}{5} = 2$$

$$\Sigma(X - \bar{X})^2 = 4+1+0+0+9 = 14.$$

7.14 ∴ Where X is 0, 1, 2, 2, 5,

$\Sigma(X - \bar{X})^2 = 14$ and s^2 =

$$s^2 = \frac{14}{4} = 3\frac{1}{2}$$

7.15 $(X - \bar{X})$ is called ..... (words)

The deviation from the mean.

7.16 State the formula, $s^2 = \dfrac{\Sigma(X - \bar{X})^2}{N - 1}$ in words.

The variance is the sum of the squares of the deviations from the mean divided by one less than the number of results.

7.17 $\sqrt{\dfrac{\Sigma(X - \bar{X})^2}{N - 1}}$ is the formula for s, which is called what?

The standard deviation.

7.18 Complete this calculation of the standard deviation from Frame 5.6 by deciding the values of a, b, c, d, s^2 and s. ($\bar{X} = 100$).

X	$(X - \bar{X})$	$(X - \bar{X})^2$
80	−20	400
90	(a)	(b)
100	0	0
110	+10	100
120	+20	400

$\Sigma(X - \bar{X})^2 = (c)$

$\frac{\Sigma(X - \bar{X})^2}{N - 1} = \frac{(c)}{(d)} = s^2 =$

$s =$

a = −10
b = 100
c = 1000
d = N − 1 = 4

$s^2 = 250$

$s = \sqrt{250}$

7.19 It is easier to calculate the range than the standard deviation. Why is the standard deviation a better measure?

It uses all the information.

7.20 What is the symbol for the standard deviation and its formula?

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N - 1}}$$

7.21 s^2 is the .....

What is its formula?

Variance.

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{N - 1}$$

7.22 If X is 1, 1, 2, 3, 3

What is the range?
What is $\bar{X}$?
What is the value of the variance?
What is the value of the standard deviation?

Range = 2
$\bar{X} = 2$

$s^2 = \frac{(1^2 + 1^2 + 0 + 1^2 + 1^2)}{4} = 1$

$s = \sqrt{1} = 1.$

7.23 If X is still 1, 1, 2, 3, 3

What is $\Sigma(X^2)$?

What is $(\Sigma X)^2$?

$\Sigma(X^2) = 24.$

$(\Sigma X)^2 = 10^2 = 100.$

7.24 $\therefore$ What does $\dfrac{\Sigma(X^2) - \dfrac{(\Sigma X)^2}{N}}{N-1}$ equal?

$24 - \dfrac{\dfrac{100}{5}}{4} = 1.$

7.25 $\therefore$ When X is 1, 1, 2, 3, 3

$\dfrac{\Sigma(X-\bar{X})^2}{N-1} = 1$ (from Frame 7.22)

and $\dfrac{\Sigma(X^2) - \dfrac{(\Sigma X)^2}{N}}{N-1} = 1$ (from Frame 7.24)

In fact, whatever the distribution *it is always arithmetically true* that

$$\frac{\Sigma(X-\bar{X})^2}{N-1} = \frac{\Sigma(X^2) - \dfrac{(\Sigma X)^2}{N}}{N-1}$$

$\therefore \sqrt{\dfrac{\Sigma(X^2) - \dfrac{(\Sigma X)^2}{N}}{N-1}}$ is another formula for calculating .....

s, the standard deviation. In fact, you all saw this to be so in the practical example in Frame 5.29.

7.26 Which of these, if any, is a correct formula for the standard deviation?

(a) $\sqrt{\dfrac{(\Sigma X - \bar{X})^2}{N-1}}$

(b) $\sqrt{\dfrac{\Sigma(X-\bar{X})^2}{N}}$

(c) $\sqrt{\dfrac{(\Sigma X)^2 - \Sigma(X^2)}{N-1}}$

contd. on opposite page

7.26 *contd.*

(d) $$\sqrt{\frac{\Sigma(X^2) - \frac{(\Sigma X)^2}{N}}{N-1}}$$

(e) $$\sqrt{\frac{\Sigma(X - \bar{X})}{N-1}}$$

(f) $$\sqrt{\frac{\Sigma(X - \bar{X})^2}{N-1}}$$

(g) $$\sqrt{\frac{\Sigma(X - \bar{X})^2}{N-1}}$$

(d) and (g)

(a) has the summation sign inside the bracket.
(b) has the wrong denominator.
(c) has both brackets in the wrong place.
(e) omits the square.
(f) omits the square root sign.

7.27 One formula for the standard deviation does not need to have the mean calculated first.
What is it?

$$\sqrt{\frac{\Sigma(X^2) - \frac{(\Sigma X)^2}{N}}{N-1}}$$

7.28 If the mean is a whole number, it is easier to use the formula

$$\sqrt{\frac{\Sigma(X - \bar{X})^2}{N-1}} \quad \text{for } s.$$

If $\bar{X}$ was calculated to be 2.3816 it would be easier to use which formula?

The one not requiring the calculation of deviations from the mean,

i.e. $s = \sqrt{\frac{\Sigma(X^2) - \frac{(\Sigma X)^2}{N}}{N-1}}$

7.29 If X is 2, 3, 3, 4, 5

$\bar{X} = 3.4$

Complete the following table:

X	$X - \bar{X}$	$(X - \bar{X})^2$	X^2
2	−1.4		4
3	−0.4	0.16	9
3	−0.4	0.16	
4			16
5	+1.6	2.56	
$\Sigma X =$	$\Sigma(X-\bar{X})=0$	$\Sigma(X-\bar{X})^2 =$	$\Sigma X^2 =$

X	$X-\bar{X}$	$(X-\bar{X})^2$	X^2
2	−1.4	1.96	4
3	−0.4	0.16	9
3	−0.4	0.16	9
4	+0.6	0.36	16
5	+1.6	2.56	25
17	0	5.20	63

7.30 $\therefore$ the variance $= \dfrac{\Sigma(X - \bar{X})^2}{N - 1}$ =

$\dfrac{5.20}{4} = 1.30$

and the variance $= \dfrac{\Sigma(X^2) - \dfrac{(\Sigma X)^2}{N}}{N - 1}$ =

The same,

$\dfrac{63 - \dfrac{17^2}{5}}{4} = 1.30$

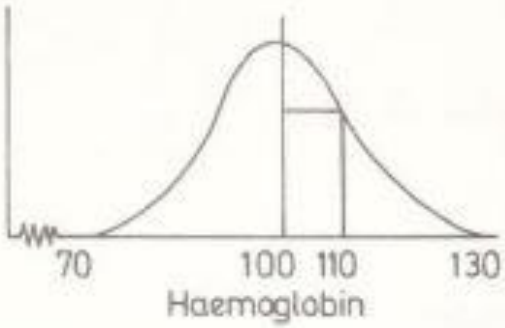
7.31 Give the formula for the standard deviation without using the mean.

$s = \sqrt{\dfrac{\Sigma(X^2) - \dfrac{(\Sigma X)^2}{N}}{N-1}}$

7.32 The standard deviation for I.Q. is about 15. What is the variance for I.Q.

About 225.

7.33 A frequency distribution in a journal looks like this. Describe it as fully as possible.



Shape - normal.
Mean/Median/Mode = 100
Standard deviation = 10
Variance = 100
Range = 60

7.34 Like the ....., which measures the middle of a distribution, the standard deviation and variance use all the data. They are more reliable measures than the ..... and ..... which measures the middle, and the ..... which measures variation.

Mean.

Median and Mode.
Range.

7.35 State 2 formulae for the variance.

$s^2 = \dfrac{\Sigma(X - \bar{X})^2}{N - 1}$

$= \dfrac{\Sigma(X^2) - \dfrac{(\Sigma X)^2}{N}}{N - 1}$

7.36 In an examination you would probably be given any formulae which are harder than these. However, the standard deviation and variance are so important that you could be expected to remember them. Can you?

Yes, I hope.

7.37 If the value of $\bar{X}$ is not a whole number, which formula would you use for the variance?

$$\frac{\Sigma(X^2) - \frac{(\Sigma X)^2}{N}}{N-1}$$

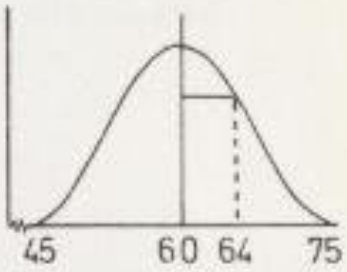
7.38 If Y is the variable rather than X, what would be the formula for the standard deviation without using Y ?

$$\frac{\Sigma(Y^2) - \frac{(\Sigma Y)^2}{N}}{N-1}$$

7.39 One measure of variation does not use all the information. It is the ....., which is used when the ..... (also unreliable) is used as the measure for the centre.

Range
Median.

7.40 Draw a normal distribution with mean 60, standard deviation 4 and range 30.



7.41 In journals you often see written: 'the mean ± the standard deviation'. For example 100 ± 15 for I.Q., seen in an article indicates what?

That the mean I.Q. is 100 and the standard deviation is 15.

7.42 **Practical Example**

Results in reality are not so arithmetically convenient as in this programme. They are usually more like those in Frames 3.1 and 3.5. Calculate the mean and standard deviation in Frame 3.1 and Frame 3.5. We will use the answers in Chapter 17 when we test to see whether babies of diabetic mothers are *significantly* bigger than those of normal mothers, so your efforts now will be put to practical use later. In fact arithmetically you could perform significance tests based on means already, but I want you to *understand* what you are doing because the tests are then much more interesting.

SUMMARY

Measures of variation are used to indicate the spread of the curve. The variance, s^2, and its square root, the standard deviation, s, are the measures of choice; the range is occasionally used.

The range of the results is easy to calculate but has the same disadvantage as the median and mode; that it is not a reliable measurement. It is used as a measure of variation when the median is used as a measure of the centre.

The range is the difference between the highest and lowest values.

The formulae to calculate the variance are

$$\frac{\Sigma(X^2) - \frac{(\Sigma X)^2}{N}}{N-1} \quad \text{or} \quad \frac{\Sigma(X-\bar{X})^2}{N-1}$$

which are numerically identical. The standard deviation in the normal distribution is the horizontal measurement from the mean to a point of inflection and equals the square root of these formulae.

Chapter 8 WHAT CORRELATION MEANS

INTRODUCTION

So far we have learned to describe qualitative data in terms of ratios and rates etc. and quantitative data in terms of the shape, middle and variation of the frequency distributions. When 2 or more different variables are measured on the same people to see whether one is associated with the other (a common practice is medical research) we describe the results in terms of correlation.

8.1 Correlation does not mean causation. The probability of getting lung cancer and the number of cigarettes smoked have been shown to be correlated. Does this mean smoking causes lung cancer?

No. No more than lung cancer causes smoking. It does show, though, that there is an association between smoking and lung cancer.

8.2 Correlation means which of the following?

(a) association.
(b) causation.
(c) living with relatives!
(d) tied together.
(e) acting the same way.

(a)

8.3 Height and weight are correlated. Increase in weight is associated with increase/decrease in height.

Increase.

8.4 When an increase in one variable is associated with an increase in another, correlation is said to be *positive*. A decrease in one variable associated with an increase in another is *negative* correlation. Size in shoe and size in hat are correlated how?

Positively.

8.5 Time spent in bed and time spent in studying are .....correlated.

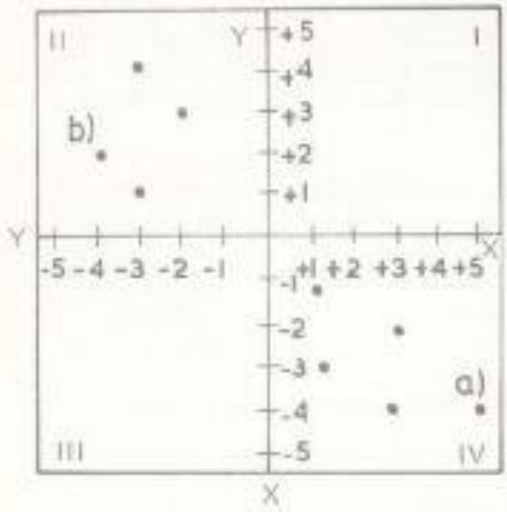
Negatively — unless you study in bed!

8.6 How are the I.Q. of parent and I.Q. of child correlated?

Positively.

8.7 Only 2 variables are considered simultaneously in terms of correlation. When this is the case the information may be represented on a 'scatter diagram'. What are the 2 variables in this scatter diagram?

X and Y.



8.8 Each dot in the last frame is where a value of X corresponds to a value of Y. At point (a) X is +5 while Y is .....?

-4

8.9 At point (b) ..... is positive and ..... is negative.

Y

X

8.10 In the quadrant marked II, X is ..... and Y is .....

-ve.

+ve.

8.11 In quadrant I, X and Y are both positive in comparison with quarter III, where X and Y are both negative. When most points lie in quadrants I and III the correlation is positive/negative.

Positive.

8.12 Which quadrants would contain most dots in negative correlation?

II and IV.

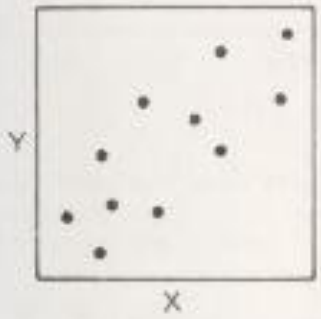
8.13 In Frame 8.7 correlation is .....

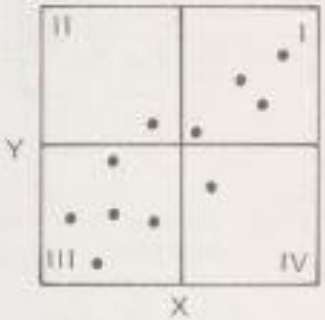
−ve.

8.14 When no correlation exists the dots in the s_____ d_____ occur roughly the same amount in all quadrants (like a non-specific rash).

Scatter diagram.

8.15 Draw the axes in this scatter diagram.

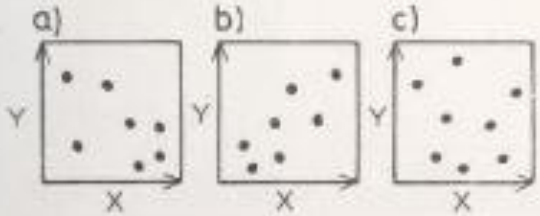




Is correlation positive or negative?

Positive.

8.16 The axes are only guide lines and need not be shown. Manage without them and complete the following:



(a) is ..... correlation

(b) is ..... correlation

(c) is ..... correlation

(a) −ve.

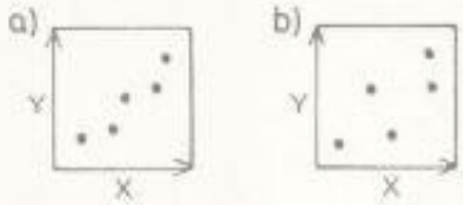
(b) +ve.

(c) no.

8.17 So far we have learned to describe the direction or the sign of the correlation. Correlation also has a descriptive tag for its size. It is *maximum* when a value of X is specific to a single value of Y (i.e. a straight line). The more the points are scattered about an imaginary line the less the correlation becomes, until when the points are scattered all over the place there is no correlation at all. It is not the slope that determines the degree of correlation but how closely the points are to a straight line.

Correlation is/is not greater in (a) than (b).

Is.

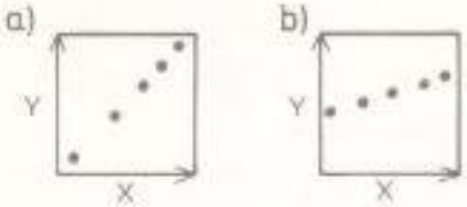


8.18 Correlation is maximum in which of the following:

(i) (a) only?
(ii) (b) only?
(iii) both (a) and (b)?

Both (a) and (b), as in both the points lie exactly on a straight line. A state of affairs very rarely met in reality.

The slope itself is what is meant by regression, but we do not discuss regression in this book.

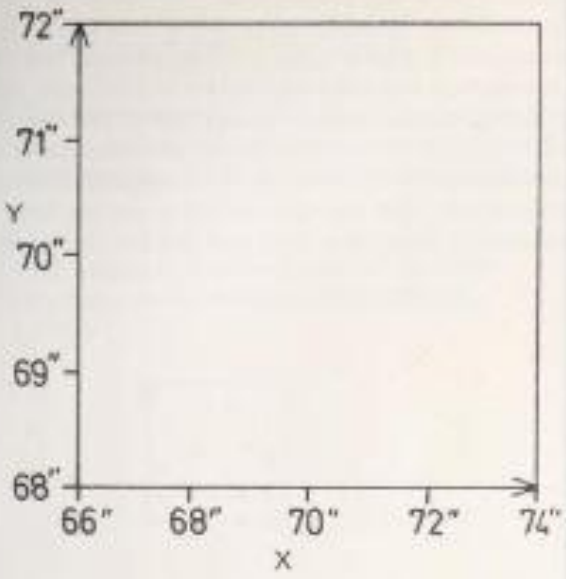


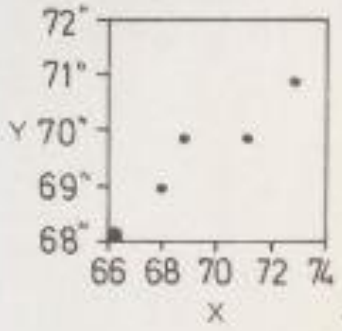
8.19 Put this information into a scatter diagram and answer the questions.

X Height of father	*Y Height of eldest adult son*
66"	68"
68"	69"
69"	70"
71"	70"
73"	71"

contd. on opposite page

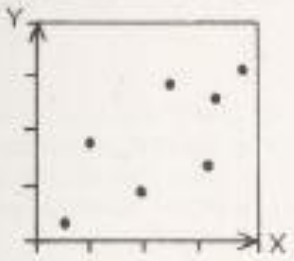
8.19 *contd.*





Is correlation present? If so, is it positive or negative?
Is it greater than below?

Yes.
Positive.
Yes.



8.20 Correlation means .....

Association.

SUMMARY

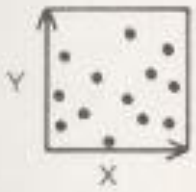
Correlation means association. It can be positive, negative or nonexistent. Positive correlation is where the variables tend to increase in size together. Where 2 variables are involved a scatter diagram may be used to represent the data. In positive correlation the dots tend to lie in the upper right-hand and lower left-hand quadrants, while in negative correlation the dots tend to lie in the other 2 quadrants. There is no preponderance of dots in any quadrant where correlation does not exist. The magnitude of correlation is indicated by how closely the dots approximate to a straight line (i.e., how narrow the scatter is about an imaginary line) and not by their slope.

Chapter 9 MEASURING CORRELATION

INTRODUCTION

You will learn here about 2 ways of measuring correlation. You would usually not be expected to remember these formulae, but having been given them you would be expected to be able to use them.

9.1 One measure of correlation is called the *Pearson correlation coefficient* and its symbol is 'r'.



You would expect 'r' to equal ..... here.

0.

9.2 How are most variables distributed?

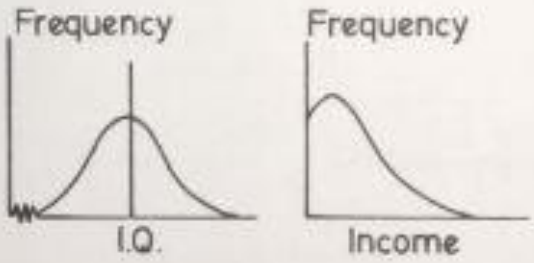
Normally.

9.3 In order to calculate 'r' *and for the value to be meaningful* both variables involved must be distributed normally. May 'r' be calculated between height and weight?

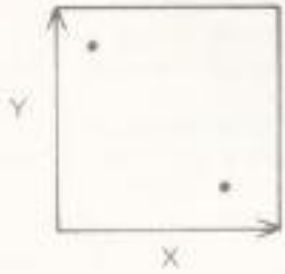
Yes. Both are distributed normally.

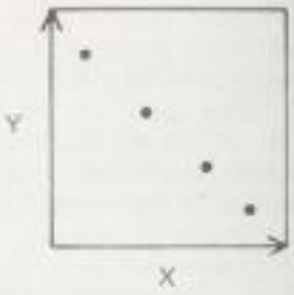
9.4 Should 'r' between I.Q. and income be calculated? Their frequency distributions are given below.

No.
Income does not follow the normal distribution shape.



9.5 Complete this diagram with r its maximum value but negative, by adding 2 dots.



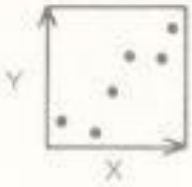


In a straight line.

9.6 r is ..... than its maximum value here and its sign is .....

less

+ve



9.7 Give a formula for s_- without looking back to Chapter 7 if possible.

$$\sqrt{\frac{\Sigma(X-\bar{X})^2}{N-1}}$$

or

$$\sqrt{\frac{\Sigma(X^2)-\frac{(\Sigma X)^2}{N}}{N-1}}$$

9.8 The denominator for r is

$$\sqrt{\frac{\Sigma(X-\bar{X})^2}{N-1}} \times \sqrt{\frac{\Sigma(Y-\bar{Y})^2}{N-1}}$$

which is what?

The standard deviation of x multiplied by the standard deviation of y.
Symbolically $s_x s_y$.

9.9 Write $s_x s_y$ without using the means.

$$\sqrt{\frac{\Sigma(X^2)-\frac{(\Sigma X)^2}{N}}{N-1}} \times \sqrt{\frac{\Sigma(Y^2)-\frac{(\Sigma Y)^2}{N}}{N-1}}$$

9.10 Is $\Sigma(X - \bar{X})^2$ the same as

$\Sigma(X - \bar{X})(X - \bar{X})$?

Yes.

9.11 The numerator for 'r' is

$$\frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{N - 1}$$

Comment.

It is very similar to the formula for the variance. It is, incidentally, called the *covariance* for X and Y.

9.12

$$\text{As } \frac{\Sigma(X - \bar{X})^2}{N - 1} = \frac{\Sigma(X^2) - \frac{(\Sigma X)^2}{N}}{N - 1}$$

Do you think

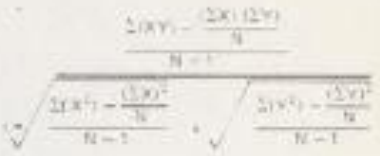
$$\frac{\Sigma(X - \bar{X})(Y - \bar{Y})}{N - 1} = \frac{\Sigma(XY) - \frac{(\Sigma X)(\Sigma Y)}{N}}{N - 1}$$

It does.

9.13 $$r = \frac{\text{covariance of X and Y}}{s_x s_y}$$

So, without using means, [as very rarely are both $\bar{X}$ and $\bar{Y}$ whole numbers]

$r = ?$



9.14 In fact the $N - 1$ term in the numerator can be cancelled with the $\sqrt{N-1} \times \sqrt{N-1}$ in the denominator, so that

$r =$

$$\frac{\Sigma(XY) - \frac{(\Sigma X)(\Sigma Y)}{N}}{\sqrt{\Sigma(X^2) - \frac{(\Sigma X)^2}{N}}\sqrt{\Sigma(Y^2) - \frac{(\Sigma Y)^2}{N}}}$$

To what does N refer?

The number of X results, or the number of Y results, i.e. the number of pairs of results (points on the scatter diagram).

9.15 This formula looks awful but we will use it now to prove that it is not too hard to use. (For your convenience it has been reproduced on a pull-out card at the back of the programme.)

To use this formula you need/need not know the means of X and Y.

Need not.

9.16 We will use the formula to calculate the value for r in this scatter diagram.

We expect here the sign of 'r' to be ..... and 'r' to equal its maximum value.

+ve.

9.17 Transfer these 3 results to complete the table below.

X	Y
0	(b)
(a)	2
4	3

N = (c)

(a) = 2.
(b) = 1.
(c) = 3, the number of pairs of results.

9.18 To use the formula for r we need to know

$\Sigma(XY)$, $\Sigma(X)$, $\Sigma(Y)$, $\Sigma(X^2)$ and $\Sigma(Y^2)$.

Complete the table on opposite page

9.18 *contd.*

(X)	(X^2)	(Y)	(Y^2)	(XY)
0	0	1	(a)	0
2	(a)	2	4	4
4	16	3	9	(c)
$\Sigma(X) = 6$	$\Sigma(X^2) = 20$	$\Sigma(Y) = (d)$	$\Sigma(Y^2) = (e)$	$\Sigma(XY) = (f)$

(a) = 4
(b) = 1
(c) = 12
(d) = 6
(e) = 14
(f) = 16

9.19 Using the formula in the pullout and the totals you have already calculated i.e.

N = 3 $\quad \Sigma(XY) = 16$
$\Sigma(X) = 6 \quad \Sigma(X^2) = 20$
$\Sigma(Y) = 6 \quad \Sigma(Y^2) = 14$

what is the value of the numerator in calculating r ?

$$16 - \frac{6 \times 6}{3}$$
$$= 16 - 12 = 4$$

9.20 What is the value of the denominator in calculating r ?

$$\sqrt{20 - \frac{6 \times 6}{3}} \times \sqrt{14 - \frac{6 \times 6}{3}}$$
$$= \sqrt{(20 - 12)} \times \sqrt{(14 - 12)}$$
$$= \sqrt{8} \times \sqrt{2} = \sqrt{16} = 4$$

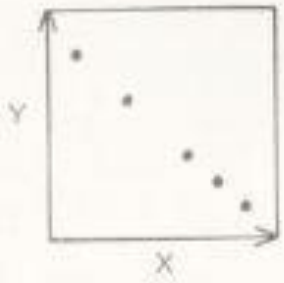
9.21 ∴ The maximum value for r =

$\frac{4}{4} = +1$ (from frames 19.19 and 19.20)

9.22 If you ever calculated r to be 5, what would be your conclusion?

You had made an arithmetical error, because 1 is its maximum value.

9.23 What is the value of r here?



−1.
The lowest possible value.

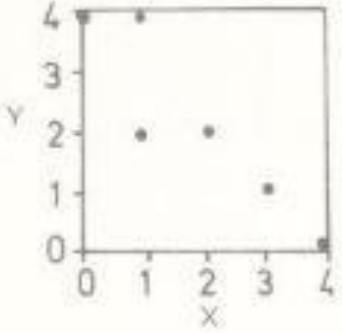
9.24 The Pearson correlation coefficient has a *sign* and a *numerical value*.

What do these signify?

The sign signifies the direction of lie of the points and the numerical value signifies how closely the points lie to a straight line.

9.25 Guess which of the following 6 values for 'r' is correct here.

−0.1, −0.5, −1.0,
−0.9, 0 +0.9



−0.9

9.26 You can use the formula to calculate 'r' in the last frame by completing this table.

(X)	(X^2)	(Y)	(Y^2)	(XY)
0		4		
1		4		
1		2		
2		2		
3		1		
4		0		
$\Sigma(X)$ =	$\Sigma(X^2)$ =	$\Sigma(Y)$ =	$\Sigma(Y^2)$ =	$\Sigma(XY)$ =

N =

$\Sigma(X) = 11$. $\Sigma(Y) = 13$.
$\Sigma(X^2) = 31$. $\Sigma(Y^2) = 41$.
$\Sigma(XY) = 13$. $N = 6$.

contd. on opposite page

9.26 *contd.*

∴ r approximately =

$$r = \frac{13 - \frac{11 \times 13}{6}}{\sqrt{\left(31 - \frac{11^2}{6}\right)\left(41 - \frac{13^2}{6}\right)}}$$

$$= 13 - \frac{143}{6}$$

$$\sqrt{(31 - 20.2)\ (41 - 28.2)}$$

$$\approx \frac{-11}{\sqrt{11} \times \sqrt{13}}$$

(≈ means approximately equals)

$$\approx \frac{\sqrt{11}}{\sqrt{13}} \approx \frac{-3.3}{3.6} \approx -.9$$

9.27 'r' is the ..... Pearson correlation coefficient.

It is used for ..... data which is Quantitative.

distributed ..... Normally.

9.28 The other coefficient of correlation is *Spearmans rank order correlation coefficient.* Its symbol is ρ (Rho.) The formula, which you need not remember, is also in the pullout. N is the same in both correlation coefficient formulae. What does it represent?

The number of pairs of results.

9.29 As the name 'rank order correlation coefficient' implies, ρ deals not with the actual results but with their rank order; if N is 6, and the best is ranked 1 or 1st, the worst would be ranked .....?

6 or 6th.

9.30 Here are the fictitious 2nd M.B. marks for 5 candidates. (b) was bottom in Anatomy and was ranked 5th or 5.

(c) is ranked ..... in anatomy.

3rd or 3.

Candidate	*Mark in Anatomy*	*Mark in Physiology*
a	60	70
b	29	60
c	51	60
d	53	35
e	45	50

9.31 Complete this table for the ranks from the last frame.

Candidate	*Rank in Anatomy*	*Rank in Physiology*
(a)	1	—
(b)	5	2½
(c)	3	2½
(d)	—	—
(e)	—	4
	Total = 15	Total = 15

1	1
5	2½
3	2½
2	5
4	4

9.32 When 2 candidates have the same score, what happens to their ranking? ((b) and (c) in Physiology in the last frame).

They each get the arithmetical average of the rankings they would have taken if there had been a slight difference between them. This keeps the totals in the ranking columns the same. e.g. if there is a joint top both rank 1½, the average of 1st and 2nd.

9.33 Had (e) got 60 instead of 50 for Physiology in Frame 9.30 what would his rank have been?

b, c and e would, had there been a slight difference, have been 2nd, 3rd, and 4th. As it is they now all rank the *average*

$$= \frac{2+3+4}{3} = 3\text{rd or } 3.$$

The sum of the ranks still equals 15. Notice that after these joint 3rd's comes the 5th.

9.34 D is the difference between a candidate's 2 rankings. N is the number of pairs of results (candidates).
Use the formula for ρ given below to complete the calculation for the 2nd M.B. results in Frame 9.30.

Candidate	*Rank in Anatomy*	*Rank in Physiology*	*D*	D^2
(a)	1	1	0	(ii)
(b)	5	2½	2½	6¼
(c)	3	2½	(i)	(iii)
(d)	2	5	−3	9
(e)	4	4	0	0
	$\Sigma = 15$	$\Sigma = 15$	$\Sigma(D) = 0$	$\Sigma(D^2)$ = (iv)

As an arithmetical check you should always find that besides both ranks summing to the same totals, $\Sigma(D)$ also equals zero.

(i) = ½

(ii) = 0

(iii) = ¼

(iv) = $\Sigma(D^2) = 15½$

N = ?

$$\rho = \frac{6\Sigma(D^2)}{N(N^2-1)}$$

$$= 1 - ?$$

=

N = 5.

$$\rho = 1 - \frac{6 \times 15½}{5 \times 24}$$

$$= 1 - 0.775$$

$$= +0.225$$

9.35 ρ is usually used if no real score can be assigned but orders of preference can be given, e.g.

2 Surgeons discuss the various operations for gall stones. They are unable to give an actual numerical value for the relative efficiency of the operations so they ..... them and calculate .....

rank ρ

9.36 These are the results. What is the value of ρ?

Operation	1st surgeon's Rank	2nd surgeon's Rank	D	D^2
A	1	3		
B	2	4		
C	3	2		
D	4	1		
	Total = 10	Total = 10	$\Sigma(D)$ =	$\Sigma(D^2)$ =

N =

ρ =

N = 4

$\Sigma(D^2) = 18$

(notice as a check $\Sigma(D)$ always = 0.)

$$\rho = 1 - \frac{6 \times 18}{4 \times 15}$$

$\rho = -0.8$

9.37 A 3rd surgeon thinks operation B is the best and that the 3 others have equal merit. What rank values would he give the operations

A = B = C = D =

(Check that the rank total is the same as for the other surgeons.)

A = 3.
B = 1.
C = 3.
D = 3.

9.38 The range of ρ is the same as r.

The maximum value of ρ is .....

With no correlation ρ is .....

Like r, ρ has a ..... and a ..... value.

+1

0

sign; numerical.

9.39 ..... is not so accurate as ..... for measuring correlation as it does not take into account the actual obtained result.

ρ; r
It is used when ranks are the only measures available or when the results are not normally distributed.

9.40 ρ is the Rank Order Correlation Coefficient described by ....., r was described by .....

Spearman.
Pearson.

9.41 Which correlation coefficient is easier to calculate?

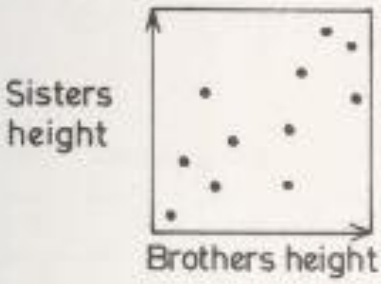
Spearman's.

9.42 To calculate these correlation coefficients you need ..... variables measured on a group of ..... people. When r is used both variables must be distributed. ....., it is the more accurate measure as the results themselves are used. Sometimes it is only possible to state a preference in which case ..... is calculated.

2
N
Normally.
ρ

9.43 Correlation coefficients of 1 rarely occur in biology and medicine. One of the nearest is that between live weight and warm dressed weight of poultry where r = +0.98.

Incidentally, the scatter diagram below represents r equal to +.6, which will help to give you some idea of the correlation coefficient size.



9.44 **Practical Example**

Below

X signifies erythrocyte sedimentation rate;
Y signifies the number of leucocytes in thousands.
Both may be thought to be distributed normally.

Using the results

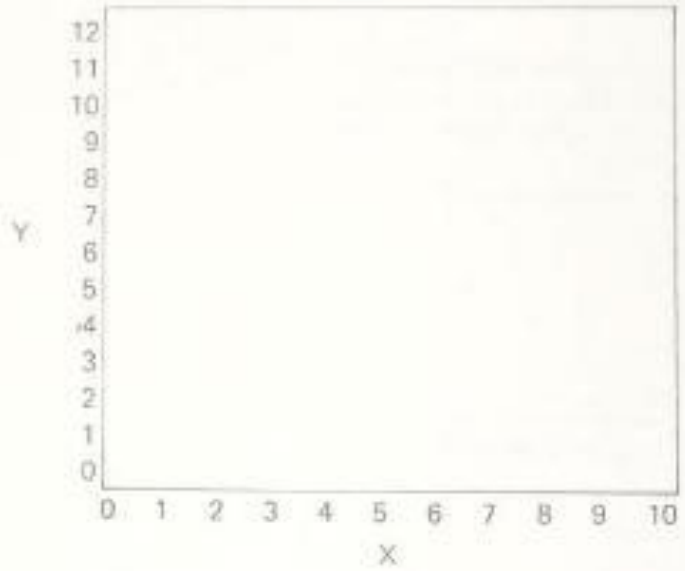
(1) Draw a scatter diagram,
(2) Guess the value of the correlation coefficient from the scatter diagram;
(3) Calculate r;
(4) Calculate ρ.

contd. overleaf

9.44 *contd.*

X	Y
1	3
2	5
3	5
5	2
5	4
5	6
7	7
7	10
9	8
10	12

(1) Scatter diagram



(2) Your estimate of the correlation coefficient =

(3) Calculation of r.

contd. on opposite page

9.44 *contd.*

X	Y	X^2	Y^2	XY
1	3			
2	5			
3	5			
5	2			
5	4			
5	6			
7	7			
7	10			
9	8			
10	12			
$\Sigma X =$	$\Sigma Y =$	$\Sigma(X^2) =$	$\Sigma(Y^2) =$	$\Sigma(XY) =$

N =

$$r = \frac{\Sigma(XY) - \frac{(\Sigma X)(\Sigma Y)}{N}}{\sqrt{\Sigma(X^2) - \frac{(\Sigma X)^2}{N}} \times \sqrt{\Sigma(Y^2) - \frac{(\Sigma Y)^2}{N}}}$$

=

(4) Calculation of ρ

X	*Y*	*Ranked X*	*Ranked Y*	*D*	D^2
1	3				
2	5				
3	5				
5	2				
5	4				
5	6				
7	7				
7	10				
9	8				
10	12				
					$\Sigma(D^2)$

contd. overleaf

9.44 *contd.*

(Check that both ranks sum to 55 and $\Sigma(D)=0$)

N =

$$\rho = 1 - \frac{6\,\Sigma(D^2)}{N(N^2-1)}$$

=

=

Does r approximately equal ρ?

SUMMARY

The 2 most frequently used correlation coefficients are Pearson's and Spearman's.

Pearson's correlation coefficient is used when both variables are normally distributed. It is symbolised by 'r' where

$$r = \frac{\Sigma(XY) - \frac{(\Sigma X)(\Sigma Y)}{N}}{\sqrt{\Sigma(X^2) - \frac{(\Sigma X)^2}{N}}\sqrt{\Sigma(Y^2) - \frac{(\Sigma Y)^2}{N}}}$$

(N is the number of pairs of results)

Spearman's correlation coefficient is used when either variable is not normally distributed, as well as when only ranks are available. It is not such an accurate measure as Pearson's and is symbolised by ρ (rho) where

$$\rho = 1 - \frac{6\,\Sigma(D^2)}{N(N^2-1)}$$

and D is the difference between rankings.

Both r and ρ range from +1 to −1. They have sign and magnitude.

+1 signifies maximum positive correlation, −1, maximum negative correlation, and 0 means no correlation at all.

Part II Ideas To Improve The Value Of Numbers

Chapter 10 POPULATIONS AND SAMPLES

INTRODUCTION

If you read some of the early volumes of the Journal of the Royal Statistical Society I think you would be surprised – maybe we would all be surprised that you were reading any more about statistics! In the 1830's research workers aimed to investigate entire populations; their task was usually impossibly hard and their work suffered accordingly. Today's research workers would consider only part, a sample, of each population and would draw their inferences from this. As you may imagine, these sample-to-population transplants are at best hazardous, with a lot of potential pitfalls – equally as hazardous, I am told, as the step from talking about love to actually proposing! In medicine it is not enough to describe a patient, you have to assess the underlying condition. In statistics it is not enough to describe the results in the sample, you have to be able to assess the worth of the particular sample.

10.1 A population is an entire group about which some specific information is required or recorded. What is the population in Frame 1.9?

All doctors in Country X.

10.2 The population is of prime importance as it is the subject of an experiment. It must be fully defined so that those to be included and excluded are clearly stated. For example in the last frame you would need to know whether those doctors are included who are retired, part-time, or on leave, or who have in fact left that country while remaining on the register. Do you?

No.

'All doctors' in that country is not a fully defined population.

10.3 In the population 'your medical faculty' you would include which of the following:

1) Teaching staff excluding part-time.
2) Teaching staff including part time.
3) The medical librarian.
4) The medical students.
5) The cleaners in the medical school.
6) The bodies in the dissecting room.

It's up to you. I would include 1) but am not very broad minded. 'Your medical faculty' is not a fully defined population until we are all perfectly clear who to include.

10.4 A statistical population need not be made up of people. We can have populations of birthweights, haemoglobin levels or blood cells so long as the population is what?

Fully defined.

10.5 A *sample* is any part of the fully defined population. A syringe-full of your blood taken now is a sample of what population?

All your blood in circulation at the moment.

10.6 Sometimes, as above, a sample is the only means we have of inferring about a population. Sampling is also slower/quicker and cheaper/dearer than the complete enumeration of the population.

Quicker; cheaper.

10.7 Any inferences from a sample refer only to the particular population defined. Among a sample of patients in your teaching hospital it is found that only patients with cancer of the lung smoke more than 40 cigarettes daily. Does this indicate that smoking more than 40 cigarettes daily is associated with cancer of the lung?

Yes. Pedantically: among the patients in your teaching hospital only.

10.8 Of course, this finding is nevertheless interesting, but only as a pointer to further research. The data on doctors in Country X tells you ..... about doctors in neighbouring countries.

Nothing.

10.9 What is a sample?

Part of a defined population.

10.10 What are $\bar{X}$ and s^2 the symbols for?

Mean and variance.

10.11 In fact $\bar{X}$ and s^2 are the symbols for the mean and variance of the sample. Guess what μ and σ^2 are the symbol for.

The mean and variance of the population.

10.12 μ and σ^2 are called mu and sigma squared. σ is pronounced ..... and represents what?

Sigma.
Standard deviation of the population.

10.13 σ is small sigma. Capital sigma is drawn ..... and means .....

Σ Add together.

10.14 A parameter is a constant used in describing a population.
σ and μ are examples of parameters and $\bar{X}$ and s are examples of statistics.
What is the difference between parameters and statistics?

Parameters refer to the population and statistics to samples from the population.

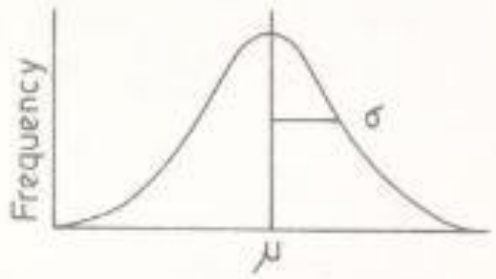
10.15 Statistics/Parameters are used to infer about Statistics/Parameters.

Statistics.
Parameters.

10.16 Each population has one/many value(s) of μ and one/many value(s) of $\bar{X}$.

One.
Many.

10.17 This is the frequency distribution of what?



A population.

10.18 If you had not enumerated a population you could estimate σ by calculating s using which formula?

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N - 1}}$$

or

$$s = \sqrt{\frac{\Sigma(X^2) - \frac{(\Sigma X)^2}{N}}{N - 1}}$$

10.19 Statistics refer to samples.
Parameters refer to populations.
How can you remember this?

S and S

P and P

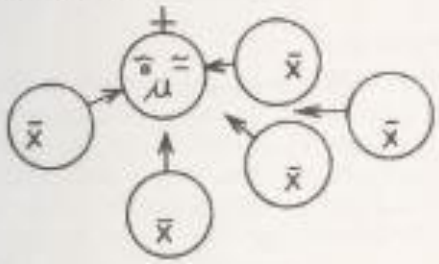
10.20 The size of the sample to be used is ideally a statistical consideration but is often limited in terms of
t - - - and c - - -

Time and Cost.

10.21 For each population there is one value for each parameter, but for each population there are many possible samples each with their own
..... estimating this parameter,
i.e. for every μ there are many possible ..... assessing it.

Statistic.
Each value of $\bar{X}$ may differ slightly but they all should approximate to μ.

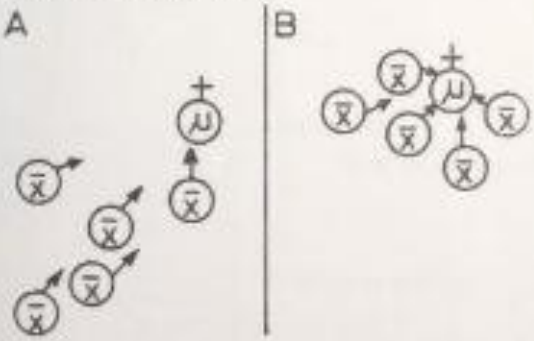
10.22 For every μ there are many possible $\bar{X}$s assessing it!



10.23 Good samples produce reliable ..... while bad samples don't.

Statistics.

10.24 The statistics shown in A are/are not better than those in B.

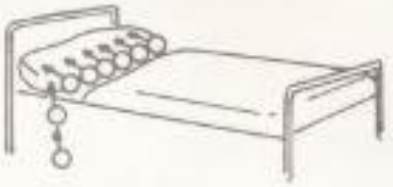


Are not.

They are not such close estimates.

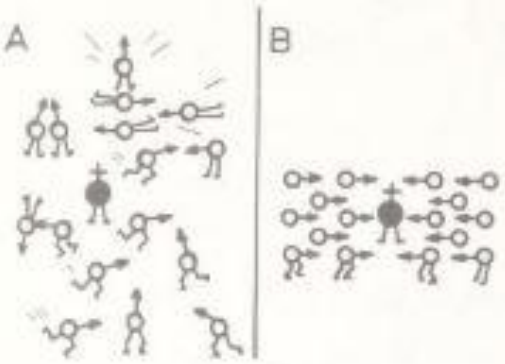
Therefore in practice we must design our methods for choosing samples so that as in B we could rely on fairly well any possible value of X.

10.25 *Precision* is a term which indicates how close statistics *are to each other.*



10.26 *Bias* is the term which refers to how far the average statistic lies *from the parameter.*

10.27 Both these ideas are important and we should aim for precise unbiased samples. (B here)



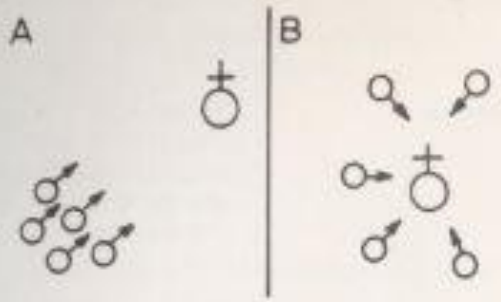
A rugby scrum!

10.28 In frame 10.24 A/B is more precise and A/B is more biased.

B
(the statistics lie closer together)

A
(the average statistic lies further from the parameter)

10.29 Here A/B is more precise but more biased.



A
The statistics lie close together but are not on the target.

10.30 Good samples have statistics which are ..... and .....

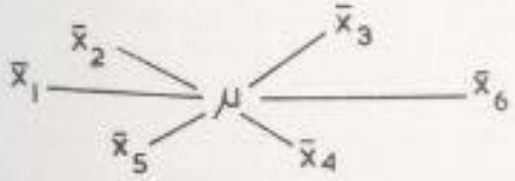
Precise Unbiased

10.31 We will discuss bias again in the next chapter. We will conclude this chapter by chatting further about precise statistics. What is precision?

How closely the statistics lie *to each other*.

10.32 Remember that precision says nothing about whether statistics are on-target. However, an unbiased experiment without precision may still easily cause a statistic to be an off-target estimate. Below is an unbiased, non-precise experiment. If your particular estimate is ..... it would not be very worthwhile.

$\bar{X}_1$ $\bar{X}_6$



10.33 The precision of a sample can be estimated.

It equals $\frac{\sqrt{N}}{\sigma}$

Here N is the size of the sample and σ is .....

The standard deviation of the population.

10.34 For any particular set of circumstances σ in the above formula is a constant.
To vary precision ..... must be varied.

N or $\sqrt{N}$

10.35 To increase precision N must be decreased/increased.

Increased.
This is what you would expect – the bigger the sample the more precise the estimate.

10.36 A sample of 100 is ..... as precise as one of 25.

Twice. $\frac{\sqrt{100}}{\sigma}$ is twice as bi as $\frac{\sqrt{25}}{\sigma}$

10.37 To double precision N must be .....

Quadrupled.
(multiplied by 4)

10.38 Research workers often ask a statistician how large a sample they need to use. The reply depends on the level of precision required and the value of .....

σ

10.39 If you know how precise an estimate you need but don't know σ you can use a pilot survey or, to be up-to-date, a mini-survey, to obtain an estimate of σ.
You would in fact calculate ..... from the pilot survey sample and substitute it in the formula;
Precision = $\frac{\sqrt{N}}{s}$

s

10.40 What is a sample?

Part of a defined population.

10.41 From any particular sample you can estimate what?

The parameters, in the relevant fully defined population only.

10.42 Give 2 characteristics of a good statistic.

It is precise and unbiased.

10.43	What is precision?	How closely the statistics lie together.
10.44	Precision is estimated using what formula?	$\frac{\sqrt{N}}{\sigma}$
10.45	N represents what?	The sample size.
10.46	To increase precision what must you do?	Take a bigger sample.
10.47	A sample which is too small may be unbiased but it is too ..... to draw valid conclusions.	Imprecise.

SUMMARY

It is not enough for you to be able to describe numbers – you must also be able to evaluate their worth when samples are used.

The population is the entire group in which you are interested. A sample is a portion of that population. The population must be clearly and exhaustively defined before a sample is drawn from it. If we are not sure whether a certain type of patient is included in the population because the population is not fully defined, we cannot be sure that any conclusion based on the sample refers to that particular type of patient. Any conclusions based on information from the sample only refer to the particular population as defined.

μ (mu) and σ (sigma) are parameters and refer to the population. $\bar{X}$ and s are the equivalent statistics in the sample and are used to estimate the parameters.

The only accurate way of estimating parameters is a complete population enumeration. Sampling is cheaper and quicker and is occasionally the only method of estimation available. The inference about a parameter using statistics is always hazardous even with good samples.

One of the characteristics of a good sample is precision or having statistics lying close together. Precision is measured by $\frac{\sqrt{N}}{\sigma}$ where N is the sample size and σ the standard deviation in the population.

σ can be estimated if unknown from a pilot survey. To increase precision the sample size must be increased. To double precision the sample size must be increased fourfold.

Chapter 11 FAIRNESS IN SAMPLING

How to be on target

INTRODUCTION

To be really 'with it', statistics must be unbiased as well as precise.

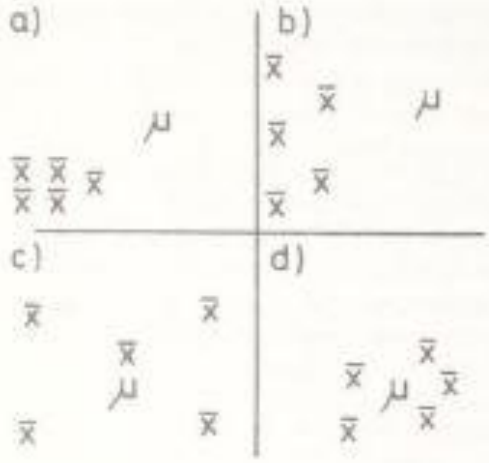
11.1 ..... is the term describing the closeness of statistics to each other.

Precision.

11.2 What is bias?

The term describing how far the average statistic is from the parameter.

11.3



Above,

..... shows good precision but is biased. (a)

..... shows good precision and is unbiased. (d)

..... shows poor precision but is unbiased. (c)

..... shows poor precision and is biased. (b)

11.4	In the last frame ..... represents the best state of affairs and ..... represents the worst.	(d) (b)
11.5	How do you improve precision?	Increase the sample size.
11.6	Before we can minimise bias we must decide how it arises. Different sorts of bias, like different diseases, require different treatments. If you wanted to find the average weight of adult males in a town you would/ would not take as your sample the weights of the rugby team? Why.	Would not. They would be a biased sample.
11.7	The commonest source of bias, as in the last frame, is in selecting the sample. The treatment for bias in sampling is *randomisation*, once described as 'the price of fairplay.' One kind of random sample is the *simple random sample*, each member of the population has an equal chance of inclusion in this type of sample. If your population is not exhaustively and clearly defined can you have a *simple random sample?*	No. If you do not know the constitution of the population the members cannot have an equal chance of inclusion. N.B. There are other kinds of random samples but we will not discuss them further in this programme.
11.8	A sample of 6 patients with disease X are required for a series of complicated tests from a population of 100 patients already numbered 00 to 99. Slips of paper numbered 00 to 99 are mixed well in a hat or sterilizing drum and the first 6 numbers drawn out are the sample. Is this a simple random sample?	Yes. All patients had an equal chance of being drawn; initially the members of the population were numbered consecutively and the required number was drawn.
11.9	Do the winning lottery tickets constitute a simple random sample?	Yes.
11.10	What is a simple random sample?	One in which each member of the population has an equal chance of selection.

11.11 What is an advantage of the simple random sample?

It guards against bias in selecting the sample.

11.12 To save time writing out numbers every time we want a random sample we can use instead published tables of random numbers. See the pull-out again. These numbers were originally chosen so that they are free from bias.
Instead of choosing 6 numbers between 00 and 99 out of the sterilizing drum, we could read off 6 numbers each with 2 digits from the table of random numbers.
Look at the table of random numbers. Start at the top left hand corner and read down the first 2 columns. Which are the 6 numbers which would constitute your random sample of patients?

Patients numbered
06
34
34
47
93
86
We will discuss the fate of the unfortunate patient No. 34 included twice, later.

11.13 Remember that before drawing a random sample we must define the population and give each member a .....

Number.

11.14 If 660 patients had constituted the population we would need to read .....columns together rather than 2.

3
Otherwise those numbered in the hundreds could not be included.

11.15 If numbers turned up which were *bigger* than the number in the population we would ignore them and continue until we had filled the sample spaces with numbers of the required size. Why is it better to number a population of 100, 00 to 99 rather than 1 to 100?

As it stands you would need to use 3 columns and would waste many of the numbers.

11.16 If the same number occurs twice in the table you include it twice in your sample, (if you are being pedantically correct). Can a person appear more than once in the sample?

Yes, in theory. Patient 34 did, back in Frame 11.12. Many people in practice would reject it a second time.

11.17 It may be a source of bias in itself if you regularly use the tables and start at the same place (or if you look at the tables while deciding where to start.) How can you avoid this?

Vary your starting point, deciding on it before seeing the tables.

11.18 The numbers are read off consecutively from your selected starting point. The numbers can be read upwards, downwards or sideways. When is this decision taken?

Before seeing the tables.

11.19 What is the variable here?

Weight gained over a 2 months period for 100 one year old children.

Weight Gained in oz. of 100 1-year-old Children in 12 months

Child's No.	*Gain*	*Child's No.*	*Gain*	*Child's No.*	*Gain*	*Child's No.*	*Gain*
00	31	25	33	50	36	75	20
01	27	26	29	51	29	76	37
02	23	27	32	52	30	77	27
03	33	28	37	53	35	78	31
04	30	29	34	54	28	79	21
05	30	30	30	55	41	80	33
06	38	31	34	56	32	81	36
07	33	32	28	57	30	82	31
08	25	33	29	58	29	83	36
09	34	34	35	59	27	84	37
10	26	35	33	60	32	85	36
11	32	36	28	61	38	86	34
12	22	37	19	62	34	87	29
13	28	38	32	63	22	88	30
14	17	39	39	64	35	89	26
15	35	40	28	65	26	90	31
16	31	41	23	66	24	91	33
17	24	42	33	67	29	92	31
18	23	43	30	68	27	93	26
19	25	44	29	69	35	94	34
20	28	45	25	70	24	95	32
21	26	46	23	71	29	96	27
22	40	47	22	72	27	97	31
23	33	48	44	73	27	98	26
24	29	49	31	74	30	99	30

11.20 In the last frame the children are numbered 00 to 99, notice, rather than 01 to 100. How many columns do you need to read together to draw a random sample using the table of random numbers?

2

11.21 We want a random sample of 10 of these children. I have decided to start in columns 7 and 8, row 3 in the table of random numbers reading sideways to the right first. What is my sample of numbers? (see the pull-out)

88 99 40
36 36 47
50 48 33
05

....

....

....

36 is included the second time.

11.22 Therefore, what is my sample of weight gains using these random numbers to choose the 10 children from Frame 11.19?

....

....

....

30 oz. 30 oz. 28 oz.
28 oz. 28 oz. 22 oz.
36 oz. 44 oz. 29 oz.
30 oz.

11.23 In my sample $\bar{X} = 30.5$ and $s = 5.8$ (compared with the population values $\mu = 30$ and $\sigma = 5$).
Do you think this is reasonably accurate?

I think so, considering the sample is fairly small.

11.24 Now, guided by the scheme below, choose your own simple random sample of 10 children.

Column No. =
Row No. =
Direction =

10 random numbers (from the pull-out) =

....

....

....

It is very unlikely that your own simple random sample will be identical to mine or anyone else's in your group. All values of $\bar{X}$ and s should approximate to 30 and 5, though, do yours?

contd. on opposite page

11.24 *contd.*

10 weight gains (Frame 11.19.) =

....

....

....

Therefore your value of $\bar{X}$ =

and your value of s =

11.25 You may be thinking that the simple random sample is a lot of bother. You have a good point. Why is it used?

It is one of the samples which protects from bias in sampling. In fact a simple random sample can occasionally be off-target but we can calculate the chances of this happening. It makes chance work for us rather than against us.

11.26 These are the steps in drawing a simple random sample, put them in order.

(a) Read off the sample of random numbers.
(b) Allot a number to each member of the population.
(c) Decide where to start in the table of random numbers.
(d) Refer the random sample of numbers to the population and read off the corresponding results.
(e) Decide in which direction to read the table of random numbers.

(b)

(c)

(e)

(a)

(d)

11.27 The reason for the random sample is not so much that it is unbiased but that it allows us to estimate the degree of bias we can expect. There is no short-cut to unbiased sampling. Some people think *'haphazard'* or *'willynilly'* samples are synonymous with the random samples. To use this method you can go through the population and choose the ones which you feel like choosing. Is this an acceptable method?

No, even though people are trying to be fair it is surprising how bias creeps in when people use haphazard sampling methods.

11.24 *contd.*

10 weight gains (Frame 11.19.) =

....

....

....

Therefore your value of $\bar{X}$ =

and your value of s =

11.25 You may be thinking that the simple random sample is a lot of bother. You have a good point. Why is it used?

It is one of the samples which protects from bias in sampling. In fact a simple random sample can occasionally be off-target but we can calculate the chances of this happening. It makes chance work for us rather than against us.

11.26 These are the steps in drawing a simple random sample, put them in order.

(a) Read off the sample of random numbers.
(b) Allot a number to each member of the population.
(c) Decide where to start in the table of random numbers.
(d) Refer the random sample of numbers to the population and read off the corresponding results.
(e) Decide in which direction to read the table of random numbers.

(b)

(c)

(e)

(a)

(d)

11.27 The reason for the random sample is not so much that it is unbiased but that it allows us to estimate the degree of bias we can expect. There is no short-cut to unbiased sampling. Some people think *'haphazard'* or *'willynilly'* samples are synonymous with the random samples. To use this method you can go through the population and choose the ones which you feel like choosing. Is this an acceptable method?

No, even though people are trying to be fair it is surprising how bias creeps in when people use haphazard sampling methods.

11.28 Of what value is a 'willynilly' sample?

Very limited.

11.29 A random sample is often beyond the reach of many practising doctors. A biopsy specimen and a syringeful of blood are not random samples but they are nevertheless useful.
A particular doctor's patients are not a random sample of the local population from which they are drawn.
What should a doctor do if he discovers an interesting fact about them?
(1) Refuse to write it up in the journals because it isn't a random sample.
(2) Write it up and call it a random sample.
(3) Write it up and point out it is a non-random sample.

(3) Then his work can be checked and followed up by others with more time or facilities.

11.30 If you read an article about an experiment in which no mention is made of whether it is in fact a random sample. What should you assume?

That it is not a random sample. If the author had gone to all the trouble of drawing a random sample he would have said so. Read it with considerable caution.

11.31 Which is better – a non-random sample labelled as such or an undefined sample?

A non-random sample. At least you know where you are.

11.32 Give 2 characteristics of a good sample.

It is precise and unbiased.

11.33 Bias in sampling can be controlled. How?

By taking, for example, a simple random sample.

11.34 What is a simple random sample?

One in which all members of the population are numbered and have an equal chance of selection.

11.35 List the steps in taking a simple random sample.

Allot numbers
Choose a starting point in the tables.
Choose the direction for reading the tables.
Read off the sample numbers.
Read off the results.

11.36 Experiments almost invariably need a control sample as a yardstick against which to measure the evidence. A control group is one identical to the experimental sample in all respects except the factor under consideration. To be unbiased the control sample is selected .....

Randomly.

11.37 X-rays of adult African males with a particular disease are being investigated. What is the control group?

X-rays of adult African males without the disease chosen at random. Sometimes in journals people omit a control group or use one which is in fact wrong in that particular experimental situation. If the experimental group is hospitalized it is wrong to compare it with people outside unless hospitalization is the factor under review.

11.38 Experiments almost invariably need some control!

11.39 Even a properly controlled experiment can have biased results, especially if these results are *subjective* (based on opinion or what a person says) rather than *objective* (based on facts or what is measured). Are the following subjective or objective?

(a) Haemoglobin levels.
(b) Patients response to a pain-killing drug.
(c) Birth weights.
(d) Number of cigarettes smoked daily.

(a) Objective.
(b) Subjective.
(c) Objective.
(d) Subjective - unless you count the stubs.

11.40 Subjective results can be very biased. Often a patient in a drug trial will sense what the doctor would like him to say. He might either deliberately try to please his doctor or displease him. Often a research doctor interprets subjective results subconsciously to fit in with his mood or theory. Does randomisation guard against this sort of bias?

No.

11.41 These sources of bias tend to be limited by using 'blind' or 'double-blind' methods. A 'blind' experiment is one where the patient does not know in which group he is, for example, whether he is receiving the drug or not (the control group is often given an inactive tablet called a 'placebo'). A 'double-blind' experiment is one where neither the patient nor the doctor is aware of the treatment received by the patient. How does this improve the bias situation?

The patients in both the control and the other group will tend to be equally misleading. The doctor in the 'double blind' situation cannot interpret the results to fit the particular theory.

11.42 How should a patient be allotted to a particular treatment?

At random.

11.43 A psychiatrist wants to see whether a new drug called 'Snuze' is effective against insomnia. His results are to be in 'number of hours slept during the first week on the drug.'

(1) How does he decide which patients receive 'Snuze'?
(1) At random among his patients suffering from insomnia.

(2) Which is the control group?
(2) Another group selected at random from his patients suffering from insomnia.

(3) The results are subjective/objective?
(3) Subjective.

(4) The control group need/need not be prescribed a placebo.
(4) Need.

11.44 Read this passage and then answer the following questions.

'An experiment is reported in a journal of physiology in which 2 different dietary regimes, A and B, were compared. Initially 120 normal and 25 underweight children were chosen and weighed clothed. For a 6 months period the underweight children were fed on dietary regime A and were given an antibiotic daily. The normal children were fed dietary regime B. At the end of the trial the mean weight gain of the underweight children was greater than the normal children.'
The samples are/are not random?

They are not mentioned as being random, so presumably they are not.

11.45 The correct/incorrect control group has been used?

Incorrect. Both groups should either be normal or underweight to begin with. Any difference could be due to this initial difference.

11.46 Why has the doctor given the antibiotic?

Even if this is ethical it is a further mistake as it adds another misleading factor. The difference could now also be due to the antibiotic.

11.47	It would/would not be better to weigh the children unclothed?	Would, because clothes vary in weight and precision would therefore be increase (because σ would be decreased.)
11.48	The results are subjective/objective?	Objective.
11.49	He therefore need/need not use a blind experiment?	Need not.
11.50	Which sample is more precise and why?	The 120 normal children, because this sample is bigger.
11.51	If you are so naive that you think this sort of article isn't published – good luck! One further consideration is important when thinking about samples. This is whether the samples are chosen before embarking on the research (*prospectively*), or whether patients already fall into the groups and it is only effects which are being compared *(retrospectively)*. Frame 11.45 is an example of a prospective/retrospective survey?	Prospective. The samples are chosen before the different diets were given.
11.52	A doctor has compared people who have had a heart attack with a control group to see whether their fat consumption has been higher. This is a prospective/retrospective study and what constitutes his control group?	Retrospective. The heart attack had already happened before the survey started. The heart attack patients were already grouped. His control group is a random sample of people identical save for the heart attack.
11.53	Retrospective studies are often so biased that they have been stigmatised as 'backward in 2 senses'. To make the experiment in the last frame a prospective study, you would do what?	Observe a group with a high fat consumption and a group with a low fat consumption and wait to contrast the rates of heart attacks.

11.54 Prospective studies are usually bigger. For example, if the average heart attack rate is 1 per 1,000 people, 100,000 would need to be observed prospectively to find 100 heart attack cases. They are also usually more costly and time consuming. What is their advantage?

They are usually less biased than their retrospective counterparts as the groups can be chosen randomly, (and moreover facilities can usually be included at the same time to investigate other relevant factors).

11.55 If you compared the I.Q.'s of people with bilharzia with others this is a prospective/retrospective study?

Retrospective. This shows another disadvantage with retrospective studies. If people with bilharzia were shown to be more stupid, you would be unable to say whether they had been stupid and contracted the disease or whether the disease had made them stupid.

11.56 What is the control group in the last frame?

The I.Q.'s of similar people without bilharzia.

11.57 Give 2 characteristics of a good sample.

It is precise and unbiased.

11.58 The size of the sample affects the precision and/and not the bias.

And not.

11.59 How can you guard against bias in sampling?

By using a random sample.

11.60 In which of the following situations is bias a problem?

(a) Unprecise
(b) Objective
(c) Retrospective
(d) Non random
(e) Subjective

(c) (d) and (e)
Retrospective
Non random
Subjective

11.61 How do you guard against bias with a subjective experiment if you can't make it objective?

Control (placebo)
Blind and double-blind experiments.

11.62 **Practical Example**

A less knowledgeable colleague than you wants to compare 2 slimming tablets. Tell him exactly what he must do to produce satisfactory results.

SUMMARY

Bias produces unreliable results because the statistics lie away from the parameter which they are to estimate – they are off-target.
Randomisation is the insurance against bias arising at the sampling stage. In fact, the simple random sample does not eliminate bias but it does allow us to estimate the size of the bias problem. To choose a simple random sample the population members are numbered and tables of random numbers are used to read off the numbers of the population members to be included in the sample. The starting point and direction for reading the tables of random numbers are chosen before the tables are opened. Numbers occurring more than once can be included more than once but higher numbers in the table than those used to number the population are ignored.
Subjective results (based on opinion rather than facts) are more prone to bias than objective results. A control group (with a placebo in drug trials) and blind or double-blind experimental designs can be used to diminish the effect of bias with subjective results. The control group must be as like the group under investigation as possible save for the variable under consideration. It should also be chosen randomly.
Retrospective (backward-looking) studies are generally more biased than prospective studies although they are usually smaller, cheaper and quicker to perform.

Part III Adapting the Numbers

Chapter 12

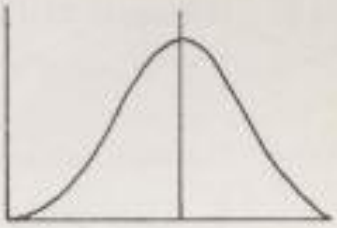
WHAT HAPPENS WHEN WE TAKE SAMPLES

INTRODUCTION

In the last chapter you all calculated a value $\bar{X}$ to estimate μ for the weight gains. Each sample was different and $\bar{X}$ varied. The way $\bar{X}$ varies is important because often you calculate only 1 value of a sample mean to estimate a population mean, and you want to know how reliable your estimate is likely to be.

12.1	For a bit of revision. What is the difference between a parameter and a statistic?	A parameter refers to a population and a statistic to a sample.
12.2	Come to that – define 'a population.'	All of something under investigation.
12.3	A sample is a ..... of the population.	Portion/part
12.4	How can you choose an unbiased sample?	Randomly.
12.5	The statistic $\bar{X}$ estimates the parameter μ. How do you make $\bar{X}$ more precise?	Increase N, the size of the sample.
12.6	Do you remember what a 'variable' is? Why can $\bar{X}$ be called a variable?	Because it varies from sample to sample.
12.7	What shape is the distribution of most variables?	Normal – bell-shaped, symmetrical with 2 points of inflection.

12.8 The distribution of $\bar{X}$ is no exception. Sketch the distribution of the sample mean.



(If the samples are large).

THE DISTRIBUTION OF THE SAMPLE MEAN

12.9 For the distribution, in fact, to be normal, the samples must be random. This also makes the samples .....

Unbiased.
This normal distribution is another virtue of the random samples. (Incidentally, if the samples are fairly large, X is distributed normally whether or not the underlying population distribution is itself normal.)

12.10 What is an unbiased sample?

One where the *average* statistic is on-target or equals the parameter.

12.11 Therefore, what is the average or mean value of the distribution of random sample means?

μ
If you got this answer right you can go straight to frame 12.15. Otherwise go to the next frame.

12.12 Imagine a population from which lots of random samples have been taken (for example the population of 100 weight gains in the last chapter),

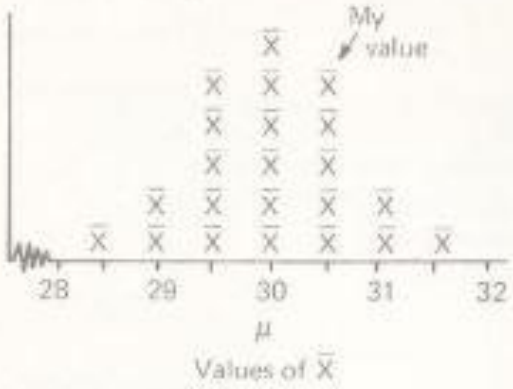
These values of $\bar{X}$ vary but are all nearly equal to .....

μ (= 30)

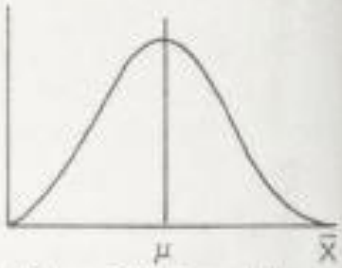
12.13 In Frame 11.24 μ was 30, your value of $\bar{X}$ was ..... and mine was 30.5

Fill in your value $\bar{X}$ here.

12.14 Therefore mark your value of $\bar{X}$ here.



12.15 Draw the distribution of the sample mean, $\bar{X}$, as fully as possible.



(The variable is usually shown at the far right end of the horizontal axis.)

12.16 What have you assumed in the last frame?

That the samples are random, and large.

12.17 Once you know the value of the standard deviation you know all you need to know about the distribution of $\bar{X}$. It is $\frac{\sigma}{\sqrt{N}}$

You have met it already (those who read this book on their heads are at an advantage!) Where?

Precision is $\frac{\sqrt{N}}{\sigma}$

It is $\frac{1}{\text{precision}}$

12.18 This seems sensible. As the precision of $\bar{X}$ increases you'd expect the variation of $\bar{X}$ to increase/diminish

Diminish.

12.19 What is the variance of the distribution of $\bar{X}$?

$\frac{\sigma^2}{N}$

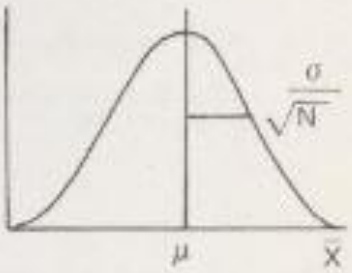
12.20 What does N represent?

The size of the sample.

12.21 If N is multiplied by 4 (i.e. you quadruple the size of your sample), the standard deviation of the distribution of $\bar{X}$ is .....and precision .....

Halved, doubled

12.22 Draw the distribution of $\bar{X}$ taken from large random samples.



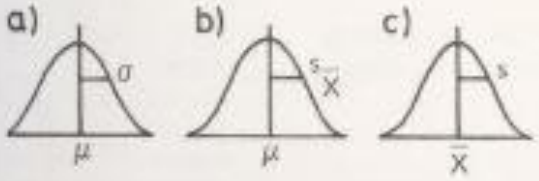
12.23 $\frac{\sigma}{\sqrt{N}}$ is so important it is given a special name and a special symbol.

$\frac{\sigma}{\sqrt{N}}$ = *the standard error of the means* = $\sigma_{\bar{X}}$ or $s_{\bar{X}}$ if σ is not known.

Write the formula for precision using this new symbol.

Precision = $\frac{1}{\sigma_{\bar{X}}}$ or $\frac{1}{s_{\bar{X}}}$

12.24



Above .....is the distribution of a population. a.

.....is the distribution of a sample from a population. c.

.....is the distribution of random sample means. b.

12.25 The value of σ for I.Q.'s of university students is 15. You each take a random sample of size 9. What is the value of $s_{\bar{X}}$

$$\frac{15}{\sqrt{9}} = 5.$$

12.26 What is $s_{\bar{X}}$ called?

The standard error of the mean.

12.27 What value had $\sigma_{\bar{X}}$ for our distribution of random samples of weight gain in Frame 11.24?

$$\frac{5}{\sqrt{10}} = 1.6.$$

12.28 If you did not know σ how would you estimate $\sigma_{\bar{X}}$?

Use s instead of σ.

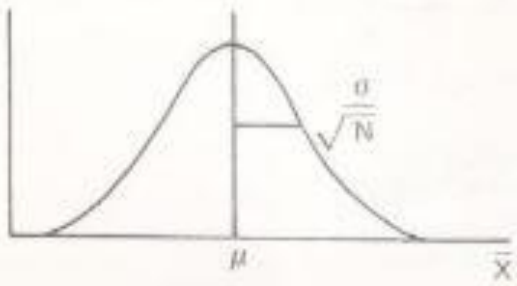
12.29 State a formula for calculating s.

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N - 1}}$$

or

$$\sqrt{\frac{\Sigma(X^2) - \frac{(\Sigma X)^2}{N}}{N - 1}}$$

12.30 What is this distribution and its standard deviation called and what exactly does it represent?



Distribution of the sample mean.
Standard error of the mean.
The distribution of all the means of large random samples of size N from a given population.

12.31 If you alter the size of your sample you do not change the value of the mean/standard deviation of the distribution of $\bar{X}$ but the mean/standard deviation of this distribution.

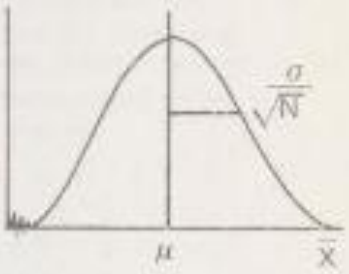
Mean.

Standard deviation.

12.32 A random sample has 2 virtues. What are they?

It is unbiased and the values of $\bar{X}$ follow a defined distribution, the distribution of the sample means.

12.33 Draw the distribution mentioned in the last frame.



Often research workers wish to compare two means. For example, soon we are going to see whether the mean birth weight of offspring of diabetic mothers differs significantly from that of normal mothers using the data from Chapter 3. Initially we assume that there is no difference, that the two means come from identical populations or the same population. To understand this fully later we will now take a quick look at

THE DISTRIBUTION OF THE DIFFERENCE BETWEEN TWO SAMPLE MEANS

12.34 From its name what do you think the variable is in the abovementioned distribution?

$(\bar{X} - \bar{X})$

say $(\bar{X}_1 - \bar{X}_2)$

12.35 You can imagine how the variable $(\bar{X}_1 - \bar{X}_2)$ is distributed.

Normally.

12.36 This is so under what conditions?

If the samples are random and large.

12.37 What was your value of $\bar{X}$ in Chapter 11, again .....?

12.38 Mine was 30.5. What value does $(\bar{X}_1 - \bar{X}_2)$ have in this case?

(Your value − 30.5)
or
(30.5 − your value)
Two of the many values of $(\bar{X}_1 - \bar{X}_2)$

12.39 If you subtracted your value $\bar{X}$ from everybody else's and everybody else did the same; all these values $(\bar{X}_1 - \bar{X}_2)$ would follow which distribution?

The distribution of the difference of two sample means. (N is small but σ is known so the distribution is normal.)

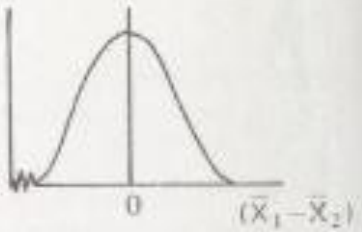
12.40 Most values of $\bar{X}$ equal or nearly equal ..... which they estimate.

μ

12.41 Some are a little bit bigger and an equal number are a little bit smaller than μ. Therefore what do you think the *average* value of $(\bar{X}_1 - \bar{X}_2)$ equals?

0
Some answers are a little bit bigger and others a little bit smaller, but the average difference is 0.

12.42 Draw the sampling distribution of the distribution of the difference between two sample means, showing the variable $(\bar{X}_1 - \bar{X}_2)$ at the far right hand side of the horizontal axis.



12.43 The standard deviation for the distribution of the mean, again, is what?

$\frac{\sigma}{\sqrt{N}}$

12.44 The variance in this distribution equals what?

$\frac{\sigma^2}{N}$

12.45 It is a fact that the variance in the distribution of the difference between two sample means is

$$\frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_2}$$

This variance in words equals what?

The sum of the variances of the two individual sample means.

12.46 i.e. The variance of the distribution

$$(\bar{X}_1 - \bar{X}_2) = \frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_2}$$

What is the standard deviation of this distribution?

$$\sqrt{\frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_2}}$$

Don't lose heart this is the programme's algebraic summit.

12.47 In our random samples from the weight gains in Frame 11.24 all values of N were equal to ..... and σ = ...

10. 5.

12.48 The distribution of $(\bar{X}_1 - \bar{X}_2)$ for these weight gains had a variance

$$\frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_2}$$

equal to what number?

$$\frac{5^2}{10} + \frac{5^2}{10} = 5$$

12.49 The standard deviation

$$\sqrt{\frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_2}}$$

equals what number?

$\sqrt{5}$

12.50 Algebraically

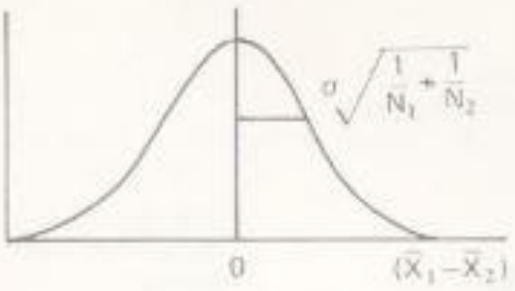
$$\sqrt{\frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_2}} = \sigma\sqrt{?}$$

$$\sqrt{\frac{\sigma^2}{N_1} + \frac{\sigma^2}{N_2}}$$

$$= \sqrt{\left(\sigma^2 \frac{1}{N_1} + \frac{1}{N_2}\right)}$$

$$= \sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

12.51 Which distribution is this?



The distribution of the difference between two sample means.

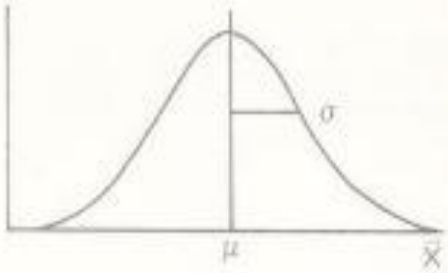
12.52 How big are the samples?

N_1 and N_2

12.53 What kind of samples?

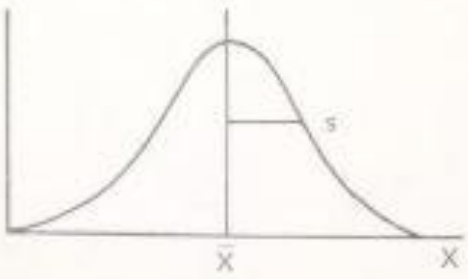
Random.

12.54 What distribution is this?



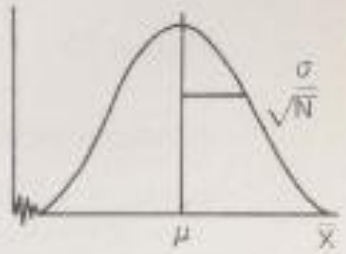
A population distribution.

12.55 What distribution is this?



A sample distribution.

12.56 Draw the distribution of the sample mean, $\bar{X}$, when the samples are random and large.



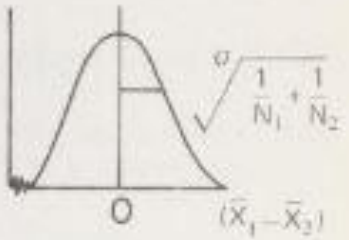
12.57 What is $\frac{\sigma}{\sqrt{N}}$ called and what is its symbol?

The standard error of the mean $\sigma_{\bar{X}}$ or $s_{\bar{X}}$

12.58 What exactly is the distribution of the sample mean?

The distribution of all sample means of random samples of size N drawn from a population.

12.59 Draw the distribution of the difference between two sample means $(\bar{X}_1 - \bar{X}_2)$



12.60 What exactly is this distribution?

The distribution of the difference between two sample means, one taken from a random sample size N_1 and one from a random sample N_2.

SUMMARY

Frames 12.54 to 12.60 serve well as the summary. You need to recognise these distributions when we use them again later.
Most readers inform me that this is the most difficult chapter in the whole book. The significance tests used in Chapters 16 and 17 to analyse actual research are based on these two sampling distributions. You will be able to perform these tests without fully understanding this chapter but of course you will be performing them rather in the dark. Should you be unhappy about the contents of this Chapter I suggest you read it again now before proceeding.
Note: To be pedantically correct sample means follow the normal curve if the samples are large and random whatever the shape of the distribution in the population. In small samples, so long as σ is known, $\bar{X}$ also follows the normal curve.

Chapter 13 THE LAWS OF CHANCE

INTRODUCTION

The last chapter was fairly difficult. However, with this chapter and the following chapter we have all the props for applying significance tests to results which are normally distributed.
This chapter itself is the basis of *all* statistical tests.

13.1 Many people talk about the likelihood, chances or odds of a particular event happening. We use the word probability. What is the probability of tossing 'tails' with a coin?

50 : 50

13.2 Probability is given the symbol 'p'. Its range is 0 to 1. When p = 0 an event is impossible. What does p = 1 mean?

That an event is inevitable.

13.3 What is p that you will die one day?

1

13.4 State an event (or maybe we should say a 'happening' in this modern age!) where p = 0.

e.g. Your swimming the Atlantic — my growing a halo, etc.

13.5 Sometimes p can be estimated logically. The probability of drawing an ace from a normal card pack is 1/13, i.e., you have 1 chance out of 13. Other times you can estimate p from the equation:

$$p = \frac{\text{total number of occurrences of the event}}{\text{total number of trials}}$$

If a surgeon transplanted 200 hearts and 1 person survived - the probability of survival here is what?

$$\frac{1 \text{ time}}{200 \text{ times}} = \frac{1}{200}$$

13.6 This means p is equivalent to

$$\frac{\text{the number of sheep}}{\text{total}}$$

which is an example of a .....

Proportion.

13.7 The probability, p, equals ..... that a coin will fall 'heads' and equals ..... that a '2' will be thrown with one of a set of dice (a die in fact!)

$\frac{1}{2}$

$\frac{1}{6}$

13.8 If one event precludes the possibility of other specified events, the events are called *mutually exclusive*. Surviving an operation, refusing an operation and succumbing during the operation are/are/are not mutually exclusive events.

Are - each of the three possibilities excludes the other two. Falling into one of the groups excludes you from the others.

13.9 Tossing a head *or* a tail with one throw of a coin are/are not mutually exclusive events.

Are. If the word or is used it infers mutually exclusive events.

13.10 Tossing a head with one coin and a tail with another coin are/are not mutually exclusive.

Are not - you can do both.

13.11 So long as events are mutually exclusive *the Addition Law of Probability* states this:
The probability that an event will occur in *one of several possible ways* is the sum of the individual probabilities of these separate events.
Therefore, what is the probability of throwing a '6' *or* a '2' with a particular one of a set of dice?

$\frac{1}{6} + \frac{1}{6} = \frac{1}{3}$

13.12 For the addition law to apply, the word *or* is seen or implied. Remember the events must be mutually exclusive. What is the probability of drawing an ace *or* a king with one cut from a pack of cards?

$\frac{1}{13} + \frac{1}{13} = \frac{2}{13}$

13.13 Is the probability of drawing an ace and a king with two cuts from the pack $\frac{2}{13}$?

No – they are not mutually exclusive events – you have two draws and can do both. The addition law does not apply.

13.14 What is the probability of tossing a 'head' *or* a 'tail' with one throw?

$\frac{1}{2} + \frac{1}{2} = 1$
i.e., it is inevitable.

13.15 When all possible outcomes of mutually exclusive events are given, their probabilities sum to one. The probability of throwing a '3' with one die + the probabilities of throwing what? = 1.

Anything but a 3, i.e. a 1 or a 2 or a 4 or a 5 or a 6 with that throw.

13.16 The probability of being blood group Rhesus +ve equals 1 minus which probability?

The probability of not being Rhesus +ve (i.e., Rhesus −ve) – its mutually exclusive event.

3.17 The probability of being Rhesus −ve is $\frac{1}{10}$
What is the probability of being Rhesus +ve.

$1 - \frac{1}{10} = \frac{9}{10}$

13.18 What are mutually exclusive events?

Those where doing one precludes any others.

13.19 What does the Addition Law of Probability state?

That with mutually exclusive events, to find the probability of one or another happening the individual probabilities are added.

13.20 What is the sum of the probabilities of a group of all possible mutually exclusive events?

1.

13.21 The probability of a pregnancy resulting in a multiple birth is $\frac{1}{80}$
What is the probability of a single birth?

$1 - \frac{1}{80} = \frac{79}{80}$

13.22 The addition law on its own is of limited use. The probability of a single birth is $\frac{79}{80}$ and of being rhesus +ve is $\frac{9}{10}$.

We cannot use the addition law to determine p for a single rhesus +ve birth. Why?

Because these events can occur together. They are not mutually exclusive.

13.23 The *Multiplication Law of Probability* applies to two or more events which do not affect each other (i.e. are independent) occurring together. Does it apply to mutually exclusive events?

No, these cannot occur together.

13.24 What are independent events?

Those which do not affect each other.

13.25 Rhesus blood grouping and multiplicity of births are independent events. The probabilities of a Rhesus +ve birth is $\frac{9}{10}$ and a single birth is $\frac{79}{80}$

What, therefore, is the probability of a pregnancy resulting in a single Rhesus +ve birth?

$\frac{9}{10} \times \frac{79}{80} = \frac{711}{800}$

(The multiplication law applies to independent events).

13.26 If the probability of a female birth is $\frac{1}{2}$ what is the probability of a female Rhesus +ve birth?
(These are also independent events)

$\frac{1}{2} \times \frac{9}{10} = \frac{9}{20}$

13.27 The probability of a birth which is female, single and rhesus +ve equals .....

(Hint: that Multiplication Law applies to 2 or more events)

$\frac{1}{2} \times \frac{79}{80} \times \frac{9}{10} = \frac{711}{1600}$

13.28 If the events involved are associated (not independent of each other), the multiplication law cannot be applied. The probability of being colourblind is $\frac{1}{12}$ and of being female is $\frac{1}{2}$.
Is the probability of being born female and colour blind $\frac{1}{24}$?

No. Colour blindness is generally associated with sex; most colour blind people are male — the multiplication law does not apply to two such associated events.

13.29 What is the probability of throwing a 6 *or* 2 with a single throw?

$\frac{1}{6} + \frac{1}{6}$
(Addition Law again)

13.30 For the multiplication law to apply the word *and* is used to connect the events.
For the addition law the word ..... connects them.

or

13.31 What is the probability of throwing a 6 with the first throw *and* a 2 with the next?

$\frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$
(Multiplication law)

13.32 What is the probability of throwing a 2 first and then a 6?

$\frac{1}{36}$ The same.

13.33 Therefore what is the probability of throwing a 6 and then a 2 *or* a 2 and then a 6.
(i.e. A 2 and a 6 either way)

$\frac{1}{36} + \frac{1}{36} = \frac{1}{18}$
(Addition Law)

13.34 Consider the same ideas in the sex of two siblings?
What is the probability of
(a) a male and then a male?
(b) a male and then a female?
(c) a female and then a male?
(d) a female and then a female?

(a) $\frac{1}{4}$
(b) $\frac{1}{4}$
(c) $\frac{1}{4}$
(d) $\frac{1}{4}$
(These are mutually exclusive and do in fact sum to 1)

13.35 Therefore what is the probability of two siblings being one male and one female (in either sequence)?

$\frac{1}{4}$ + $\frac{1}{4}$ = $\frac{1}{2}$
male then female — female then male

13.36 Complete this table for two offspring

Sequence	*Family*	*Probability*	
M M	2 males	$\frac{1}{4}$	
F M	1 male and	$\frac{1}{4}$	= $\frac{1}{2}$
M F	1 female	$\frac{1}{4}$	
F F	2 females	?	

$\frac{1}{4}$

13.37 Notice that the probability of one female and one male is twice that for two males. This is because two males can only arise in one sequence but one of each sex can arise in .....

Two.
If events can occur in more than one sequence the overall probability is the sum of the probabilities for each sequence.

13.38 7 males in a family only arise in one sequence: male and male and male etc. What is the probability of 7 offspring being all male?

$\frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2}$
$= (\frac{1}{2})^7 = \frac{1}{128}$

13.39 6 males and 1 female can arise in ..... sequences.

7.

i.e. M M M M M M F or
M M M M M F M or
M M M M F M M or
M M M F M M M or
M M F M M M M or
M F M M M M M or
F M M M M M M

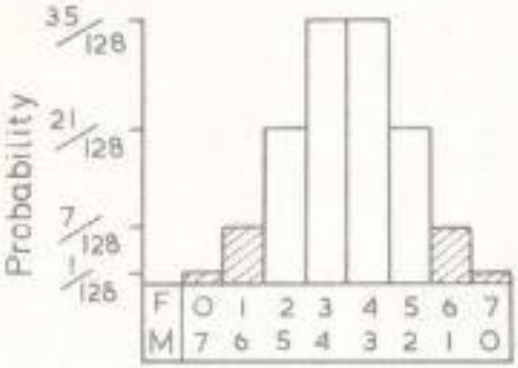
Each of them has a probability $(\frac{1}{2})^7$ so that the overall probability is what?

$(\frac{1}{2})^7 + (\frac{1}{2})^7 + (\frac{1}{2})^7$ + etc.
$= 7 \times (\frac{1}{2})^7 = \frac{7}{128}$

13.40 The number of sequences in which 2 females can arise is 21. The probability for 2 females and 5 males in any order is $\frac{21}{128}$.

Similarly the probability of 3 females = $\frac{35}{128}$ and so on.

These probabilities are represented in the diagram.



All these alternatives are mutually exclusive.
What is the total probability? 1

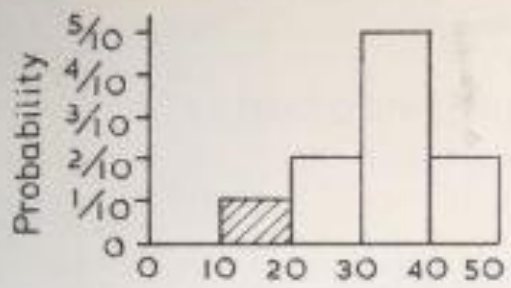
13.41 Probability can be represented by size of area so long as the total possible area equals ..... and the area drawn to scale. 1.

13.42 What is the probability of 7 children consisting of 0 or 1 females *or* 0 or 1 males.
(The shaded area in Frame 13.40)

$$\frac{1}{128} + \frac{7}{128} + \frac{7}{128} + \frac{1}{128}$$

$$= \frac{16}{128} = \frac{1}{8}$$

13.43

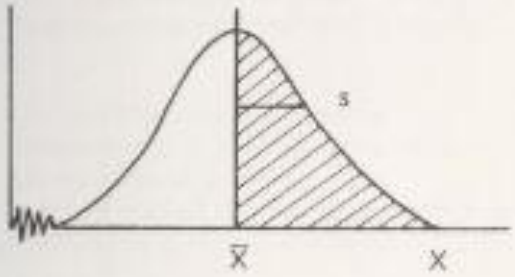


What is the probability of having a value less than 20 in this histogram? (Also the shaded area)

$\frac{1}{10}$

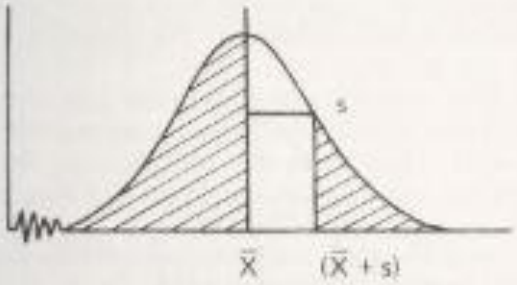
(the area is 1/10 of the whole)

13.44 What is the probability of a result greater than X in this sample distribution?



$\frac{1}{2}$

13.45 What is the probability of a result falling below X or above (X + s)?



$\frac{1}{2} + \frac{1}{6} = \frac{2}{3}$

13.46 What does probability mean?

Likelihood, chances or odds.

13.47 If the chances are 50 : 50

p = ...

$\frac{1}{2}$

13.48 Give the formula for estimating p.

Total number of occurrence / Total number of trials

13.49 When would you add probabilities together?

With mutually exclusive events when wanting the probability of one event or another.

13.50 When would you multiply probabilities?

When the probability of two or more events occurring together is required and they are not associated.

13.51 When events can occur in more than one sequence the overall probability is the .....

Sum of the probabilities for the individual sequences.

SUMMARY

Probability means likelihood, chances or odds. It has the symbol 'p' and ranges from 0 (impossibility) to 1 (inevitability). It is estimated from the formula:

$$\frac{\text{total number of occurrences of the event}}{\text{total number of trials}}$$

The *Addition Law of Probability* applies to mutually exclusive events which are events such that the occurrence of one event excludes the possibility of any other taking place. The probability of one *or* more mutually exclusive event is the sum of the individual probabilities. All the probabilities of mutually exclusive events sum to one.

When the word *and* replaces *or* we use the *Multiplication Law of Probability* such that the probability of two or more events occurring together (e.g. Event A and Event B) is the product of their individual probabilities. This is only true if the probabilities are not associated in any way, *i.e.* if they are independent.

When events can occur in more than one sequence the overall probability is the sum of the probabilities for the individual sequences.

Where the total area is one unit, proportional area can be used to represent probability.

Chapter 14

STANDARDISING THE NORMAL CURVE

INTRODUCTION

In this chapter we learn to apply the ideas about probability to the normal distribution. This is the last problem before going on to using numbers to answer questions.

14.1 What is the most widely occurring frequency distribution in medicine?

The normal.

14.2 You are to learn about a new characteristic feature of the normal curve. What characteristics do you already know?

It is symmetrical and bell-shaped with two points of inflection.

14.3 The new characteristic is about probability, p. Is probability, a ratio, a proportion or a rate?

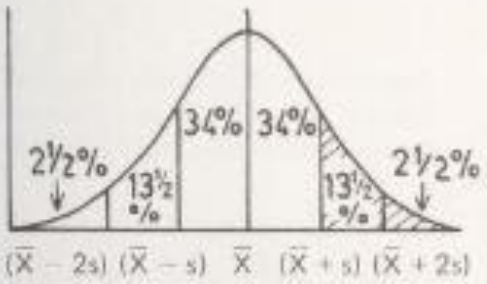
A proportion.

$$\frac{\text{Number of occurrences of an event}}{\text{Total number of trials}}$$

14.4 A percentage is 100 times a ratio/rate/proportion?

Proportion.

14.5 What is the *percentage* area beyond $(\bar{X} + s)$ here?



16%

14.6 What is the proportion of area beyond $(\bar{X} + s)$ in the last frame?

0.16

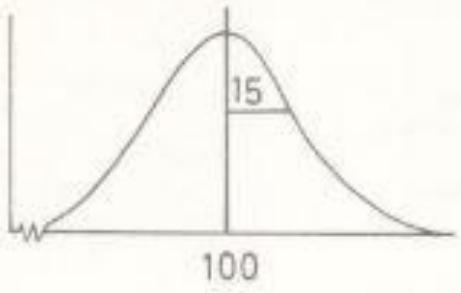
14.7 Therefore, what is the *probability* of a result being bigger than $(\bar{X} + s)$ in that normal curve?

0.16

14.8 These probabilities and % areas apply to all normal curves.
What is the probability of obtaining a result lower than two standard deviations below the mean
(i.e. below $(\bar{X} - 2s)$ in Frame 14.5)

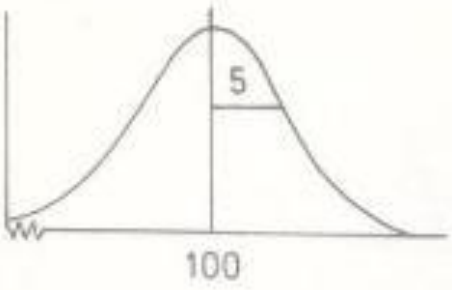
$p = 0.025$ (2½%)
The same as that above $(\bar{X} + 2s)$

14.9 Here is a curve representing I.Q.'s.
What is the probability of having an I.Q. above 130?
(Refer to Frame 14.5 if you need to)



130 = $(\bar{X} + 2s)$. The probability equals that above $(\bar{X} + 2s)$
$p = 0.025$.

14.10 This curve represents haemoglobin.
What is the probability of having a haemoglobin level above 110?

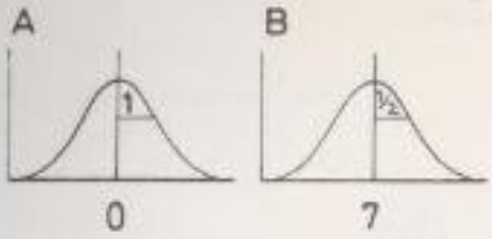


0.025, the same.

14.11 You are as unlikely to have an I.Q. of 130 or bigger as to have a haemoglobin level of 110 or above. Why is this?

Because both values lie 2 standard deviations beyond the mean on a normal curve.

14.12 Here the probability beyond 7½ on Graph B is the same as beyond ..... on Graph A.



1.

14.13 On Graph A in the last frame the value 2 lies ..... standard deviation (s) above the mean, and −1 lies ..... standard deviation(s) below the mean.

2.
1.

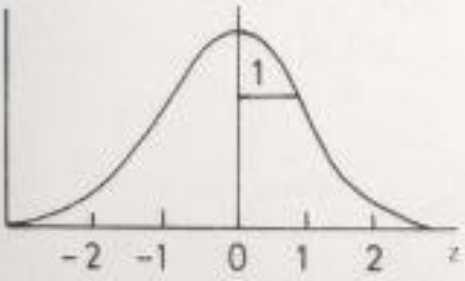
14.14 In fact in Graph A in Frame 14.12 the numerical value itself tells you the number of standard deviations you are from the mean.
−3 on this normal curve is where?

3 standard deviations below the mean.

14.15 Because Graph A in Frame 14.12 is easy to use it is given a special name — *the standard normal curve*
It has a mean value .....
It has a standard deviation .....
and variance .....

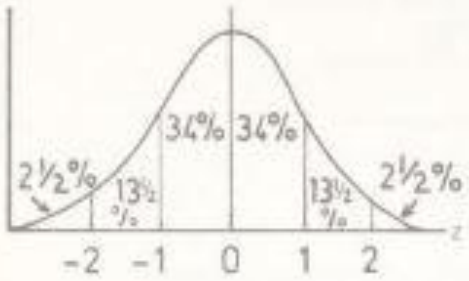
0.
1.
1.

14.16 Below is a drawing of the standard normal curve. The variable is given the symbol ..... and represents what?



z.
The number of standard deviations from the mean on a normal curve.

14.17 What is the probability of a z value less than -1 on the standard normal curve?
(The diagram from Frame 14.5 is repeated here after modification).



0.16

14.18 A z value of -1 on the standard normal curve is the same as the value (.....) on all normal curves.

$(\bar{X} - s)$.

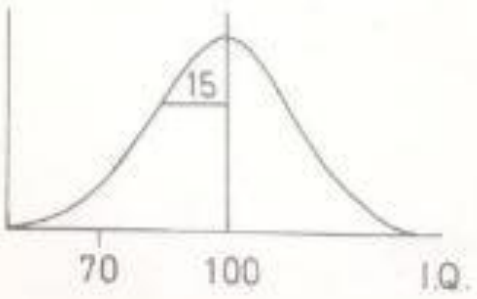
14.19 Conversely the result 8% on normal curve B in Frame 14.12 is equivalent to ..... on the standard normal curve.

+3.
It is 3 standard deviations above the mean.

14.20 To recap, the z value on the standard normal curves equals .....
.....
.....
on all normal curves.

The number of standard deviations above or below the mean.

14.21 With the result 70 here, z = ...



−2.

14.22 By calculating z as in the last frame every value on any normal curve can be related to the ..... normal curve.

Standard.

14.23 The z value equals the number of standard deviations from the ..... on any normal curve.

Mean.

This equals the

$$\frac{\text{distance from the mean}}{\text{the standard deviation}}$$

i.e. $z = \dfrac{\text{the particular result} - \text{the mean}}{?}$

The standard deviation.

14.24 Using this formula, the value 4 on the normal curve with mean 10 and standard deviation 6 is equivalent to a z value of what?

$\dfrac{4-10}{6} = -1.$

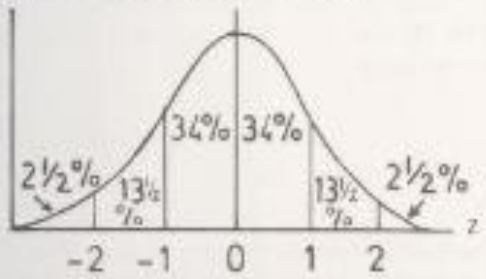
14.25 This means it lies ..... standard deviation(s) ..... the mean.

1.

below.

14.26 Therefore the probability of obtaining a result equal to or lower than 4 on a normal curve with mean 10 and standard deviation 6 is what?

0.16



14.27 Give the formula for calculating z.

$z =$

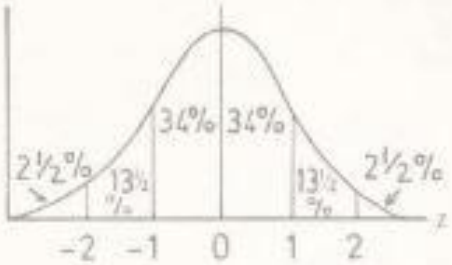
$z = \dfrac{\text{the result} - \text{the mean}}{\text{the standard deviation}}$

14.28 The formula in the last frame for z can/cannot be used for all normal distributions.

Can. This is the purpose behind the standard normal curve and z.

14.29 This diagram relates z to .....

The probability p, of obtaining a higher or lower value of z.



14.30 Therefore any value on any normal curve can be equated to ..... by first calculating .....

p.

z.

14.31 Usually, in using the normal curve to answer questions we are interested, at the same time, in values bigger than $+z$ or smaller than $-z$.
In Frame 14.29 the probability p of a result bigger than $z = +2$ or smaller than $z = -2$ is .....

0.05 or 5%
(i.e. 2½% + 2½%)

14.32 Instead of using the standard normal curve itself we can construct tables of z and the equivalent p values.
Using the diagram in Frame 14.29 and the result from the last frame we have

p.	0.32	?
z.	1	2

? = 0.05

14.33 Do you understand where the 0.32 came from in the last frame where $z = 1$?

Yes, I hope. 16% of results are beyond $z = +1$ and 16% below $z = -1$ giving a total of 32%, *i.e.* p = 0.32.
(The z value and p value refer to both ends of the normal curve together.)

4.34 Useful values for z (N.B. plus or minus) and the equivalent p values are given below:

p	0.10	0.05	0.02	0.01
z	1.6	2.0	2.3	2.6

Explain again what z = 2.0, p = 0.05 means.

The probability of a result bigger than z = +2 or smaller than z = −2 is 0.05.

4.35 z is the variable on which curve?

The standard normal curve.

4.36 z represents .....
.....
and may be calculated using the formula
z = ?

The number of standard deviations from the mean.

$$z = \frac{\text{the result} - \text{the mean}}{\text{the standard deviation}}$$

4.37 Use it, and the table in Frame 14.34 which has been transferred to the pull-out at the back of the book, to answer the following question.
What is the probability of a result bigger than 22.9 or smaller than 9.1 in a normal distribution of mean 16 and standard deviation 3 ?

$$z = \frac{22.9 - 16}{3} \text{ or } \frac{9.1 - 16}{3}$$

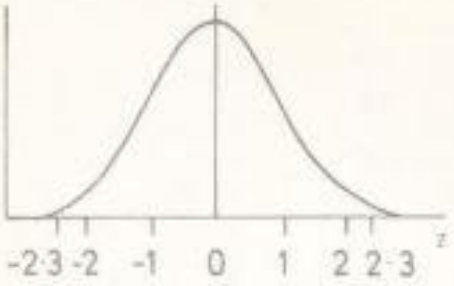
= +2.3 or −2.3

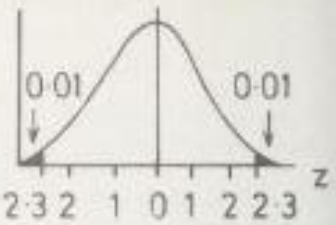
From the table, p is 0.02.

4.38 In the table, p includes both ends of the curve. If you had only been interested in the last frame in the probability of a result bigger than 22.9 you would have ..... the p value given in the table.

Halved.

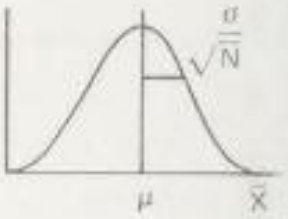
14.39 Complete the diagram below to indicate from the table $z = 2.3$ where $p = 0.02$





Total shaded area = 0·02

14.40 Draw the distribution of the sample mean! (From memory, I hope)



14.41 For any particular result $\bar{X}$ from this normal curve,

$z =$
(using the formula)

$$= \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}}$$

the result − the mean / the standard deviation

Use this in the next frame.

14.42 Assume that psychology students have an average I.Q. of 120 with a standard deviation 15. You perform I.Q. tests on a random sample of 36 such students and calculate $\bar{X}$ to be 125. What is z's value and the probability of such an $\bar{X}$ value as 125 or bigger turning up?

$\bar{X}$ =
μ =
σ =
N =
z =

$\bar{X} = 125$
$\mu = 120$
$\sigma = 15$
N = 36
$z = \frac{125 - 120}{\frac{15}{\sqrt{36}}} = +2$

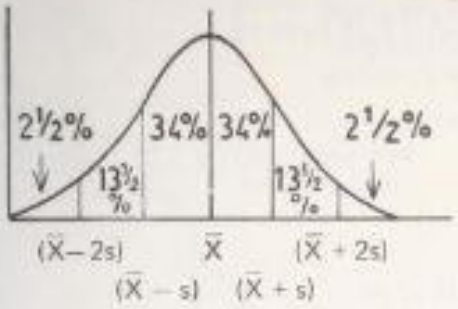
p = (from the tables)
but, p = (here)

The equivalent value of $p = 0.05$ from the tables. We only want the 'or bigger' end and therefore the required probability = ½ of .05 = .025

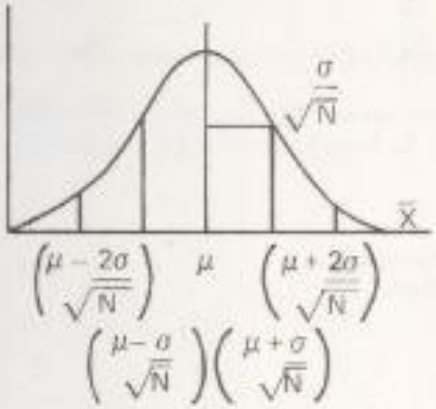
14.43 95% of the results lie between ..... and .....here.

$(\bar{X} + 2s)$

$(\bar{X} - 2s)$



14.44 Look at the distribution of the sample mean below and compare it with the last frame.



95% of sample means lie between ..... and .....

$\left(\mu + \frac{2\sigma}{\sqrt{N}}\right)$ and $\left(\mu - \frac{2\sigma}{\sqrt{N}}\right)$

14.45 We can be 95% certain that a value of $\bar{X}$ is within ..... of μ.

$\frac{2\sigma}{\sqrt{N}}$

14.46 Sometimes μ is unknown and we require to estimate it using $\bar{X}$. We are 95% confident that our particular $\bar{X}$ is within ..... of μ.

$\frac{2\sigma}{\sqrt{N}}$

14.47 $\left(\bar{X} + \frac{2\sigma}{\sqrt{N}}\right)$ and $\left(\bar{X} - \frac{2\sigma}{\sqrt{N}}\right)$

are called the 95% *confidence intervals* for estimating μ. If σ is unknown the 95% confidence intervals for μ may be written ..... and .....

$\left(\bar{X} + \frac{2s}{\sqrt{N}}\right)$ and $\left(\bar{X} - \frac{2s}{\sqrt{N}}\right)$

Provided of course the sample is random and in fact the sample is fairly large.

14.48 We wish to estimate μ for I.Q. of all university students. We take a random sample of 100 students and find the mean result in this sample to be 115 with standard deviation 10. We are 95% confident that μ is between

$$\bar{X} + \frac{2s}{\sqrt{N}} \quad \text{and} \quad \bar{X} - \frac{2s}{\sqrt{N}}$$

Which equals ..... and ..... here.

$115 + \frac{2 \times 10}{\sqrt{100}}$ and

$115 - \frac{2 \times 10}{\sqrt{100}}$

= 117 and 113.

14.49 The distribution of the sample mean is only distributed normally if the samples are.....

Random.

14.50 The 95% confidence intervals can only be used to estimate μ if the sample is ..... and .....

Random and fairly large. (Say more than 30 or σ known).

14.51 ..... and ..... are the 95% confidence intervals for μ if the sample is random.

i.e. $\left(\bar{X} + \frac{2\sigma}{\sqrt{N}}\right)$ and $\left(\bar{X} - \frac{2\sigma}{\sqrt{N}}\right)$

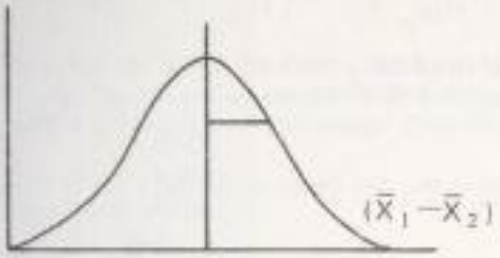
14.52 When are these 95% confidence intervals used?

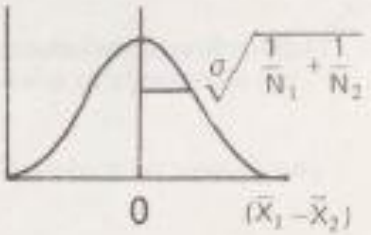
When we require to estimate a parameter (e.g. μ) from a statistic ($\bar{X}$)

14.53 We calculate the mean birth-weight of a random sample of 36 children of diabetic mothers to be 110 oz. with a standard deviation of 30. What can we say about the mean birth-weight of all such babies?

We are 95% confident it lies between $110 + \frac{60}{6}$ and $110 - \frac{60}{6}$
= 120 oz. and 100 oz.

14.54 Complete again the distribution of the difference of 2 sample means





14.55 You have a particular value of $(\bar{X}_1 - \bar{X}_2)$

Which formula would you use to calculate the equivalent value of z?

$$z = \frac{\text{the result} - \text{the mean}}{\text{the standard deviation}}$$

$$z = \frac{(\bar{X}_1 - X_2) - 0}{\sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

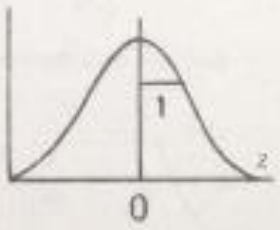
$$= \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

14.56 $z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}}$ and $z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$

Will be used again soon.
You do not need to remember them, although I hope you could derive them for yourself. Could you?

? Yes.

14.57 Sketch the standard normal curve.



14.58 z = the number of what?

Standard deviations from the mean.

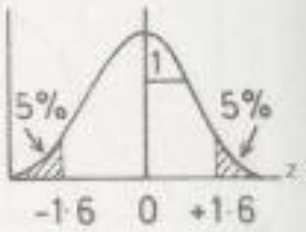
14.59 What is the importance of the standard normal curve?

The z value can be derived for all results from normal curves.

14.60 Using what formula for z?

$$z = \frac{\text{the result} - \text{the mean}}{\text{the standard deviation}}$$

14.61 Sketch the value of p used in tables relating p to z, where z equals 1.6.



14.62 If you are interested in only one end of the curve how can you use the table?

$\frac{1}{2}$ the recorded value of p.

SUMMARY

All normal curves can be adjusted so that the probability of obtaining certain results or bigger can be calculated. They can be adjusted to the standard normal curve which has mean equal to 0 and standard deviation 1. The result of the standard normal curve, z, equals the number of standard deviations a result on any other normal curve lies from the mean. z can be calculated using the formula:—

$$z = \frac{\text{the result} - \text{the mean}}{\text{the standard deviation}}$$

Therefore for the distribution of the sample mean

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}}$$

and the distribution of the difference between two sample means.

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

Tables relating p to z are readily available. The p tabulated value is the probability of obtaining a result bigger than +z or less than −z added together. If you only want the probability at one end of the curve the tabulated p value is halved.

A confidence interval can be used to predict a likely range of values for a parameter using statistics. The most commonly used confidence intervals are to predict μ from $\bar{X}$. The 95% limits are in large random samples.

$$\left(\bar{X} + \frac{2s}{\sqrt{N}}\right) \text{ and } \left(\bar{X} - \frac{2s}{\sqrt{N}}\right)$$

By substituting results from a random sample in these formulae we have a range within which we are 95% confident μ falls on condition that the sample is fairly large (say bigger than 30.)

Well done. From now on we will use only the ideas which we have already met.

Part IV Using Numbers to Answer Questions

Chapter 15

IDEAS BEHIND SIGNIFICANCE TESTS

Firstly, a short story, which is partly true.

A physician and a surgeon traditionally went to play golf on Thursday afternoons (time and weather permitting). At the Nineteenth hole it was decided that they should each toss a coin. Should both toss 'heads' or both toss 'tails' they would re-toss their coins until the unfortunate threw a 'head' (and bought the drinks) and the other threw a very profitable 'tail'.

On the first three occasions the physician tossed 'tails' and the surgeon 'heads'. Very willingly did the surgeon, although a Scot, buy the beverages. On the fourth successive occasion the surgeon emptied his pockets rather less willingly.

Returning home after the fifth successive Thursday's expenditure, the surgeon muttered to his wife: 'Och' (for he was Scottish) 'I'm sure *the physician is above board*, but I do think there is something rather uncanny in his "tail tossing" ability'. However, the surgeon's wife was able to re-assure her husband: 'It is obviously just bad luck on your part *due entirely to chance* – ignore it'.

The sixth week the surgeon tossed the unlucky 'head' yet again. Although feeling rather anti-physicians, he didn't comment. After the seventh game the surgeon's wife was faced with a very belligerent husband (surgeons can be belligerent!). She agreed that enough was enough. Even though there was no proof that the physician was employing a trick (for these results could be entirely due to chance) it was very suspicious. *'The line must be drawn somewhere'* she said, 'If it happens again, you must play golf with somebody else'.

Ideas like these are very commonly used in analysing experimental data, although the circumstances are usually rather different. The story has some statistical morals.

1. Any set of results involving data subject to chance variation, could be *'due entirely to chance'* as the surgeon's wife pointed out.
Statistics can never *prove* anything.
2. Statisticians assume that chance is the only factor initially. Like the surgeon, they give the benefit of the doubt – they assume *'the physician is above-board'*. The surgeon's wife initially thought that the results were *'due entirely to chance, too.'*
3. In statistical significance tests, as in the story, the time comes when *'the line must be drawn somewhere'*. Otherwise, no conclusions can be drawn – no action taken. Chance could always be to blame, but the time comes when the evidence is such that it is more realistic to assume some other factor.

INTRODUCTION

In earlier times medical progress, when it occurred, tended to leap forward, e.g. the discovery of the use of B_{12} and insulin. The results were obvious. However, in the last decade many advances have been made by a series of research workers each contributing a small improvement to the overall picture. On such occasions the only way to decide that an improvement really exists is by careful experimental design and analysis.
In this chapter we consider the ideas behind the general format of all significance tests.

15.1 Scientific research usually follows these steps:
a) *observation* of a phenomenon.
b) *postulation* of a theory to account for the observation.
c) *prediction* of a result on the basis of the theory.
d) *experiment* designed to test the prediction.
e) *analysis* of experimental results.
f) *conclusion* as to whether or not to accept the theory.
With which of these steps is statistics involved.

d, e and f.

15.2 Sometimes people say that the very nature of particularly medical data with its inherent variability makes its scientific analysis impracticable. This is rubbish – statistics depends on variability for its very existence. If all patients reacted the same way we could use simple arithmetic. However, statistics can only supply a measure of doubt, not .....

proof.

15.3 As the surgeon's wife in our story said, the results could always be due to what?

Entirely to chance.

15.4 Statistically, like the surgeon's wife, we always initially assume that the results obtained are *'due entirely to chance'* variation. This is the same as the legal approach, a kind of initial guilt/innocence.

Innocence.

15.5 The idea in italics in the last frame is rather pedantically called the Null Hypothesis. Look at Frame 11.44 again. What was the Null Hypothesis?

That any difference between these regimes as indicated by the results, was only due to chance.

15.6 The *alternative* to the Null Hypothesis is that the results obtained indicate that there is a situation which is more than we can reasonably account for by chance. What is the Null Hypothesis?

That the results are only due to chance.

15.7 *'The line must be drawn somewhere.'* Persistently saying that the Null Hypothesis could always be true, gets us nowhere. The time must come when the evidence is such that we must stop supporting the Null Hypothesis and give our allegiance to the .....

Alternative.

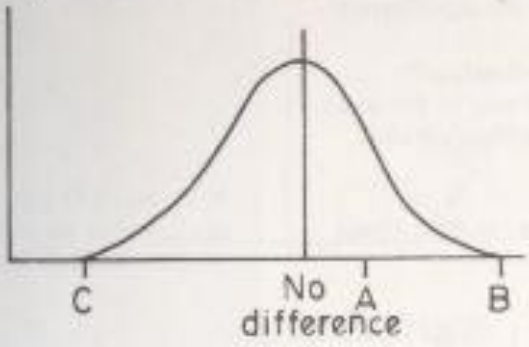
15.8 When this point is reached our conclusion is that we now accept/reject the Null Hypothesis and accept/reject the alternative theory.

Reject.
Accept.

15.9 Dr D is conducting a drug trial to decide whether a particular type of pneumonia responds better to injections of:
a) longacting penicillin, b) crystalline penicillin.
What is the Null Hypothesis and its alternative?

The Null Hypothesis is that any difference is just due to chance variation. The alternative is that the difference in response is more than can be expected by chance.

5.10 Initially we assume the Null Hypothesis is true.
Does the value A, here, support the Null Hypothesis more than the value B?



Yes. It is nearer the centre where no difference is specified.

5.11 In the last frame B could still be due to chance variation (the Null Hypothesis) but it is more likely due to .....

The alternative that there is a real difference.

5.12 In Frame 15.10

With the result A we would ..... the Null Hypothesis.

Accept.

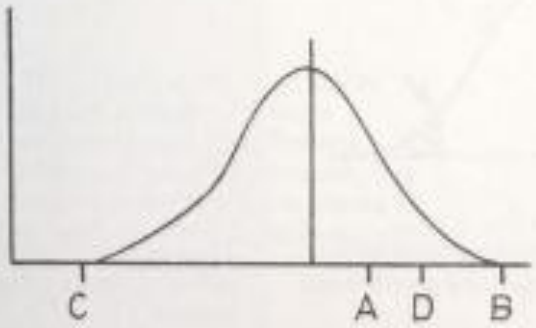
With result B we would ..... the Null Hypothesis.

Reject.

With result C we would ..... the Null Hypothesis.

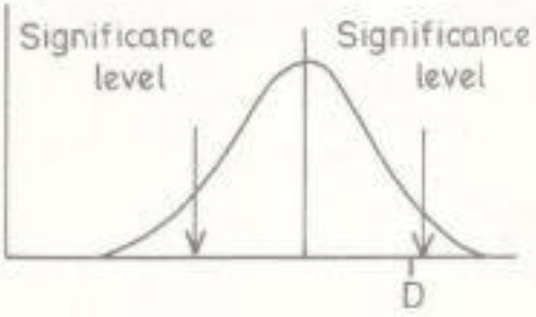
Reject.

5.13 With result D here would you;
a) reject the Null Hypothesis
b) accept the Null Hypothesis
c) not know what to do?



You say:
a) You are easily convinced
b) You are a doubting Thomas
c) You must make up your mind.

15.14 We must make up our minds and draw some conclusion with the point D. As the surgeon's wife said in the story 'The line must be drawn somewhere'. The actual line is called *the significance level*.
We reject the Null Hypothesis with extreme results at either end of the scale. Therefore we need a significance level at both ends.
Here they are.
What conclusion may be reached about D now?

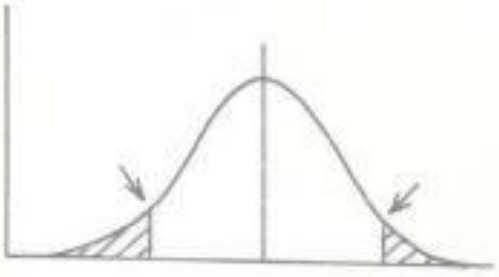


With result D and this significance level we still accept the Null Hypothesis.

15.15 Why is a significance level necessary?

It enables decisions to be made.

15.16 You could choose any significance level you like but one commonly used, the .05 level shown below is such that the *total* probability of a more extreme result at *either* end is .05.
The shaded area under this curve represents how much of the total?

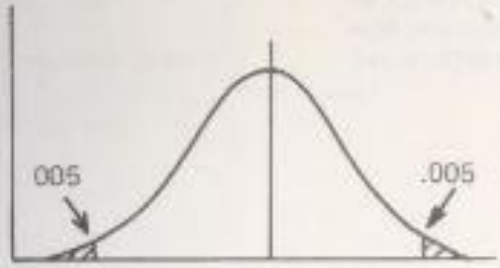


The total shaded area is .05 or 5% as the .05 significance level includes the probability at either end.

15.17 What does the .05 significance level mean in the case where 100 trials are performed?

That in 5 of these trials a more extreme result would occur by chance.

15.18 The other commonly chosen level is called the ..... significance level shown below.



.01
(.005 at each end)
occasionally the .001 level is used.

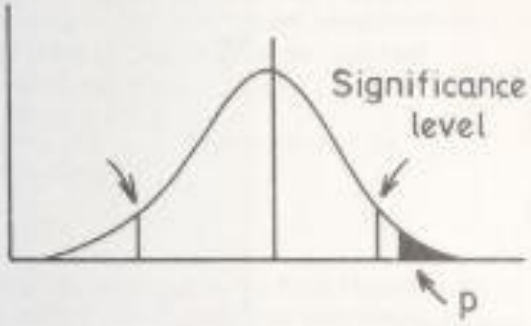
15.19 The ..... is the yardstick against which the evidence in support of and against the Null Hypothesis is measured.

Significance level.

15.20 What is the .05 significance level?

The line at which the probability of a more extreme result is .05.

15.21 The amount of evidence in *support* of the Null Hypothesis is called p. If p is less than the significance level you accept/reject the Null Hypothesis.



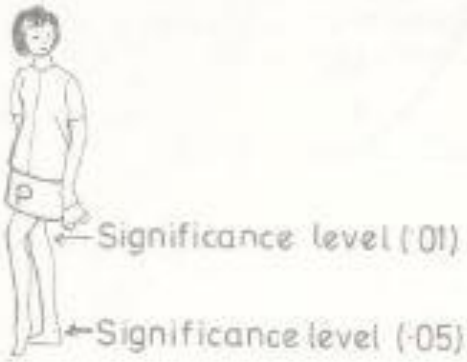
Reject.
There is insufficient evidence to support the Null Hypothesis.

15.22 The value p is the probability of such a result or a more extreme result than the one obtained, arising by chance. The smaller p the less the evidence to support the Null Hypothesis and the greater the evidence to support the .....

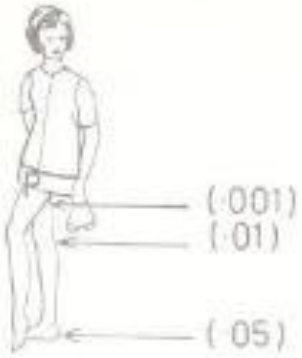
Alternative.

15.23 Imagine p represents a mini skirt. The smaller p the more mini the skirt! When p becomes less than the significance levels there is insufficient evidence to support the Null Hypothesis and now we can subscribe to a real difference!

Vive la difference!



15.24 As p becomes smaller and smaller and creeps beyond the very small significance levels so the difference becomes more and more apparent. (Statistically significant)



15.25 If p is less than .01 it is more/less significant than if it was only less than .05.

More.

15.26 This symbol, $>$ means 'bigger than.' At the end of an article in a journal you read '$.05 > p > .01$'. Therefore what conclusion is drawn at the .05 significance level?

p is less than .05. The Null Hypothesis is rejected and you accept the real difference alternative. (The skirt is midi.)

5.27 '.05> p> .01'.
At the .01 significance level your conclusion would be .....

That you still accept the Null Hypothesis, the evidence has not yet crept beyond this significance level.

5.28 You read '.01> p'. What is your conclusion?

You reject the Null Hypothesis and accept the alternative at this significance level.

5.29 If you calculate p to be .04, relative to the .05 level, .05 ..... p
(Symbol)
and your conclusion is?

.05 > p.
You reject the Null Hypothesis and accept the alternative theory.

5.30 These are the steps in performing a significant test. Number them in the correct order.
a) Calculate p.
b) State the Null Hypothesis and its alternative.
c) Draw conclusions.

a) is 2nd
b) is 1st
c) is 3rd.

5.31 Dr C wonders whether more boys or more girls get a particular complication of bilharzia. Of the 7 cases reported 6 were male and 1 female.
(Fictitious Data)
State the Null Hypothesis and its alternative.

The Null Hypothesis is that any difference in the results is due to chance.
The alternative is that there is a real difference between boys and girls!

5.32 Initially we assume the Null Hypothesis is correct. Therefore, the probability of a boy suffering the complication rather than a girl by chance is .....

$\frac{1}{2}$

5.33 Look at Frame 13.40 again. What is the probability of the 7 cases being 6 boys and 1 girl?

$\frac{7}{128}$

15.34 The p value is the value of such a result or a more extreme result occurring by chance (including both ends of the scale). From Frame 13.40 p equals .....
(The shaded areas)

$$p = \frac{1}{128} + \frac{7}{128} + \frac{7}{128} + \frac{1}{2}$$

$$= \frac{16}{128} = 0.125$$

(7 of one sex is more extreme than 6 to 1, and both ends are included.)

15.35 p' = 0.125
Your conclusion?
(Your p value is greater than the significance levels).

You accept the Null Hypothesis at all significance levels because $p > .05$ and $p > .01$.

15.36 You have just performed your first significance test.
List the stages.

a) State the Null Hypothesis and alternative.
b) Calculate p.
c) Draw the conclusion.

15.37 Assume all 7 cases had been males.
By completing the 3 stages, perform the significance test again.
a) The Null Hypothesis and alternative remain the same.
b) Your p value (Frame 13.40).
c) Your conclusion is .....

$$p = \frac{1}{128} + \frac{1}{128}$$

$$= 0.016$$

$$.05 > p > .01$$

p is less than .05 – You accept the alternative at this level.

p is more than .01 – Here still accept the Null Hypothesis

15.38 The different significance levels have resulted in a different conclusion. Less evidence is required to accept a theory at the .05 significance level than at the .01. This is the reason why a theory accepted at .01 is said to be more 'significant'. It is also the reason why if you accept the Null Hypothesis it can be due to one of two causes.

Either: there is insufficient evidence as yet to accept the alternative

or

.....

There is in fact no real difference.

5.39 To distinguish between these two causes you must do what to your samples?

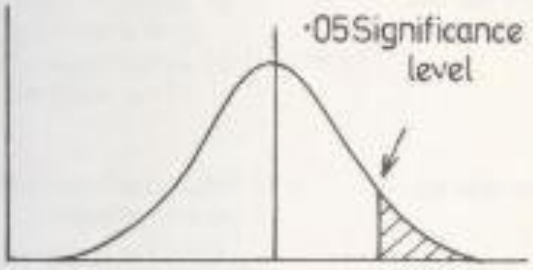
Increase their size.

5.40 Occasions arise when the theory does not involve both ends of the scale, but only one. A new drug may be more expensive and unless it is better than the old, people are not interested. They are not interested in which outcome?

Whether the old drug is better than the new. (Unless the new drug is better than the old one you can forget this other extreme.)

5.41 When only one end is important the significance levels used are still .05 and .01, but the probability areas now only apply to one end. *i.e.* for the one-tailed-test here the shaded area is ..... of the whole area.

.05 or 5%.



5.42 When the significance level only refers to 1 tail, of course, the p value you calculate also only applies to that tail. If $p > .05$, your conclusion, as before, is .....

That you accept the Null Hypothesis – either there is no real difference or there is insufficient evidence.

5.43 Another time when only one tail is used is when previous knowledge can exclude one possible extreme. Mrs H's theory is that bilharzia lowers the I.Q. of the patient (she knows it doesn't increase it). The stated ..... enables you to decide whether both or only one tail should be used.

Alternative to the Null Hypothesis.

15.44 Frame 11.44 required one/both tail(s). Why?

Both. The theory relates to either of the dietary regimes being better. With one end we would only be interested in one regime being better.

15.45 A drug company runs a trial of a new drug Y and its older counterpart X, initially on hamsters, to see whether Y is better than X. Six hamsters responded better to Y and one to X. (Each hamster received each drug but with a sufficiently long time interval in between so that there was no carry over of the effect of the drug). This requires use of one/both tail(s).

One. We are only concerned with the new drug being better than the old.

15.46 To draw conclusions from the last frame we must follow what steps?

a) State the Null Hypothesis and alternative.
b) Calculate p.
c) Draw conclusion.

15.47 State the Null Hypothesis and its alternative.

That any difference is due to chance.
That Y is better than X.

15.48 What is the probability of a hamster improving more with drug Y than drug X, assuming the Null Hypothesis is correct.

$\frac{1}{2}$

15.49 To calculate p here you do/do not include both ends.

Do not

15.50 Using Frame 13.40 again, p here =

$$\frac{1}{128} + \frac{7}{128} = \frac{8}{128} \simeq .06$$

Remember p is the probability of that result or a more extreme result occurring by chance but only at one end here.

15.51 Symbolise this result using the significance levels, p, and the 'greater than' sign (>).

$p > .05$

15.52 Therefore what conclusion do you draw?

Either there is no difference or there is insufficient evidence.

15.53 To distinguish between these two conclusions in the last frame you must do what?

Increase the sample size.

15.54 What is the first stage in performing a significance test?

State the Null Hypothesis and its alternative.

15.55 What decision is based on this alternative?

Whether one or both ends of the scale are to be used.

15.56 What does the Null Hypothesis state?

That any difference is due to chance.

15.57 What is the purpose of a significance level?

It enables decisions to be made.

15.58 Which significance levels are commonly used?

.05 and .01 (occasionally .001)

15.59 A significance level of ..... enables more alternatives to the Null Hypothesis to be accepted. However with it the Null Hypothesis will be wrongly rejected in ..... cases out of 100.

.05

5 (This is a price you must pay to reach a conclusion)

15.60 Your next step in performing a significance test is to calculate p. What does p represent?

The probability of such an extreme or more extreme result occurring by chance, assuming the Null Hypothesis is correct.

15.61 If p is less than the significance level, what is your conclusion? (*e.g.* .01 > p.)

The Null Hypothesis is rejected and the alternative accepted.

15.62 Otherwise if p > .05, what is your conclusion?

Either there is insufficient evidence to reject the Null Hypothesis or there is no real difference.

15.63 **Practical Example**

Look back at the golfing story, where the surgeon was left puzzling about the physician's ability to toss 'tails' so consistently. After how many rounds should the surgeon's wife have persuaded her husband to call a halt? Remember, if they both tossed heads or both tossed tails they threw again. Ignore the probability of this happening and just calculate, the Null Hypothesis being assumed correct, the probability of the surgeon throwing heads and the physician tails and vice versa.

SUMMARY

Statistics deals with material subject to inherent variability and helps by providing a measure of doubt about theories. These theories can never be proved. Significance tests enable research workers to draw conclusions.

The stages in performing a significance test are:

a) State the Null Hypothesis and its alternative.
b) Calculate p.
c) Draw conclusions.

The Null Hypothesis states that the experimental results are not due to the theory, but only due to chance variation. It is accepted as true until sufficient evidence is collected to reject it.

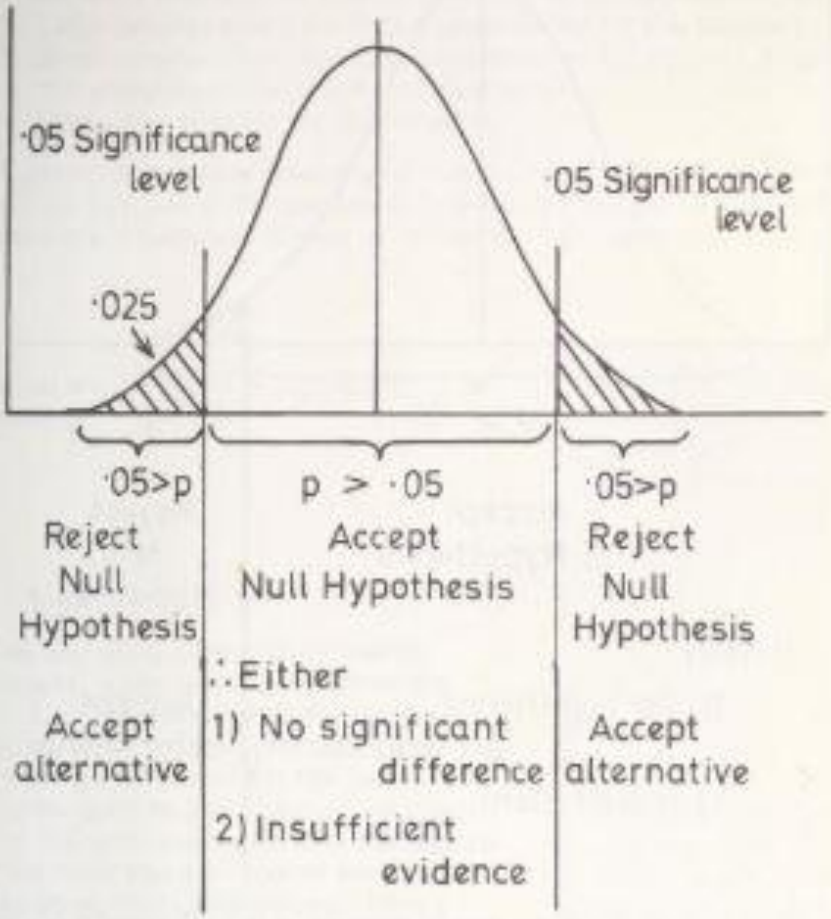
The usual significance levels are .05 and .01. They prove a yardstick against which the evidence is measured. The .05 significance level means that in 5 times out of 100 (a probability of .05) such an extreme value or more extreme value would occur by chance. The Null Hypothesis is rejected more often if the .05 level is used rather than the .01 level and more significant differences are found, but the Null Hypothesis is wrongly rejected in 5 tests out of 100.

The p value for the experimental result is the probability of the actual experimental result or more extreme results arising by chance alone. Usually p includes the equally extreme results at both ends of the scale (a two-tailed test). In this case the significance levels also include both ends. If one end of

SUMMARY (contd.)

Two examples of conclusions:

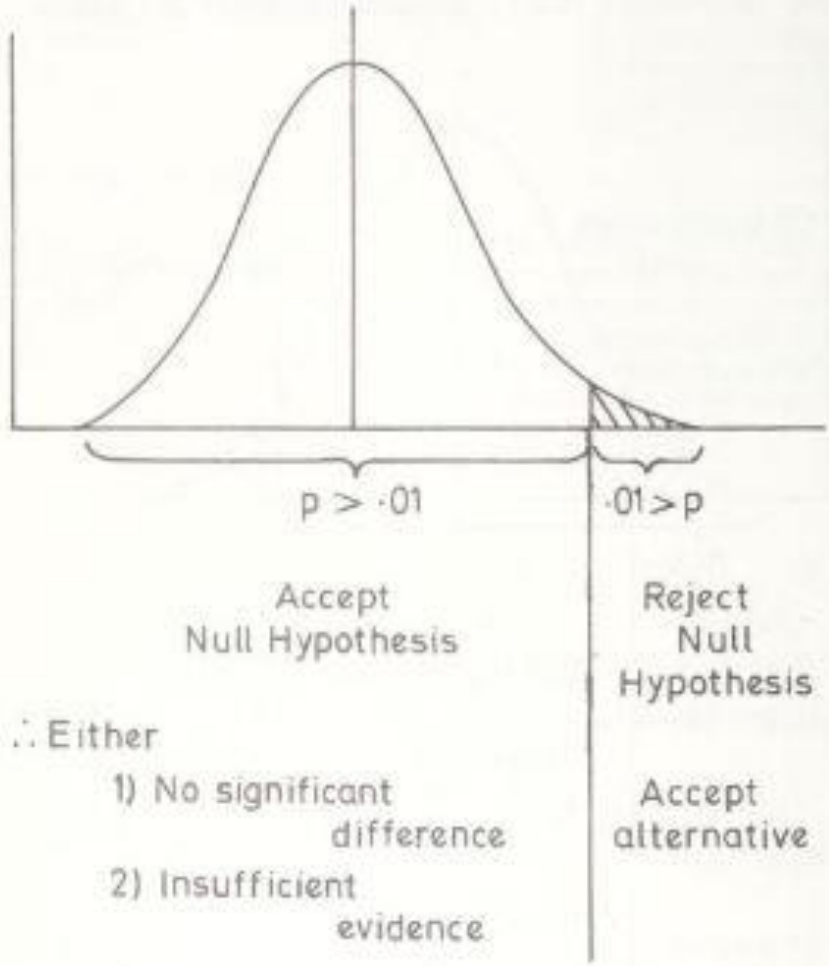
A. 2 TAILED TEST. 05 SIGNIFICANCE LEVEL



contd. overleaf

SUMMARY (contd.)

B. 1 TAILED TEST, ·01 SIGNIFICANCE LEVEL



If .05 > p > .01 the alternative is accepted at the .05 significance level, but not at the .01 level. An alternative to the Null Hypothesis accepted at .01 is more 'significant' than one at .05. Occasionally results are significant at the .001 significant level which is very significant. Even when this is so it does not *prove* that there is a real effect. It means that we should provisionally accept the idea of a real difference rather than suppose that a very improbable chance result has occurred.

Chapter 16 SIMPLE TESTS WITH 'z'

INTRODUCTION

The ideas of significance tests can be extended to various practical situations. These last 4 chapters will show you how to apply significance tests to:—

1. Large samples where the data is quantitative (in this chapter).
2. Small samples where the data is quantitative (in the next chapter).
3. The correlation coefficients (in chapter 18).
4. Qualitative data (in the final chapter).

Further ideas used in this chapter involve the distribution of $\bar{X}$ and $(\bar{X}_1 - \bar{X}_2)$, and z. Re-read the summaries of chapters 12 and 14 if your memory is rusty and if need be re-read the two chapters.

1 What are the stages in significance testing?

State the Null Hypothesis and its alternative.
Calculate p.
Draw conclusions.

2 We are going to test these results. Mrs H. wants to know whether the I.Q. of children with bilharzia is lower than normal. She uses an intelligence test which has been so designed to give a population mean of 100 with a standard deviation 12. She finds that her random sample of 36 students with bilharzia have a mean result of 96.
State the Null Hypothesis and its alternative.

The Null Hypothesis is that any experimental difference is only due to chance variation, and the alternative is that children with bilharzia have a lower I.Q. than normal.

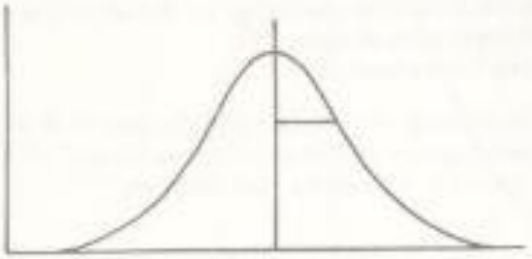
This test involves ..... tail.

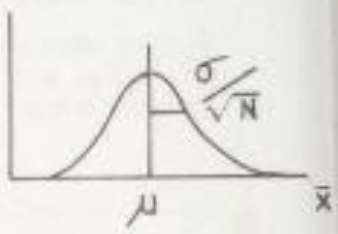
One.
We are not testing whether they they have a higher I.Q.

3 We initially assume that the .....
..... is true.

Null Hypothesis.

16.4 This being so, we can initially assume that children with bilharzia are no different from the general population, so far as their I.Q. is concerned. Therefore, Mrs H's value $\bar{X}$ can be assumed to have been taken at random from the general distribution of $\bar{X}$. Complete the general distribution of $\bar{X}$, the sample mean.





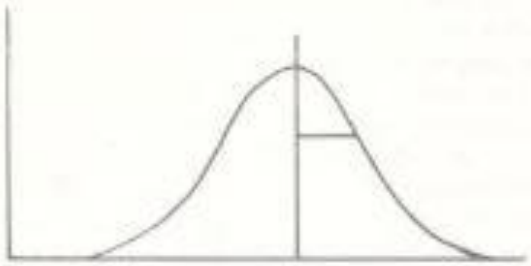
16.5 What is the other name and symbol for its standard deviation?

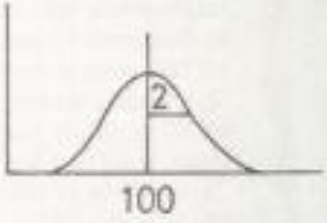
The standard error of the mean, $s_{\bar{X}}$

16.6 Draw the specific distribution of $\bar{X}$ drawn from the population in Frame 16.2 using the values for μ, σ and N given there.

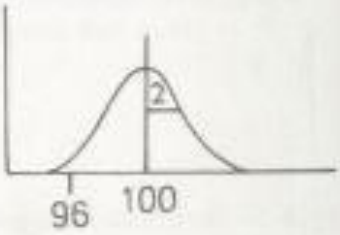
$\mu = 100$, $\sigma = 12$, $N = 36$

$\therefore \frac{\sigma}{\sqrt{N}} = \frac{12}{\sqrt{36}} = 2$





16.7 Mark in Mrs H's result $\bar{X}$.
How many standard errors away from the mean is it?
What is its z value?



(2 below −2.)

6.8 We can check that $z = -2$.
Complete this formula for calculating z.

$$z = \frac{\text{the result} - \text{(a)}}{\text{(b)}}$$

(a) The mean.
(b) The standard deviation.

6.9 For the distribution of $\bar{X}$ in general

The mean = ?

The standard deviation = ?

The mean = μ

The standard deviation

$$= \frac{\sigma}{\sqrt{N}}$$

z =

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}}$$

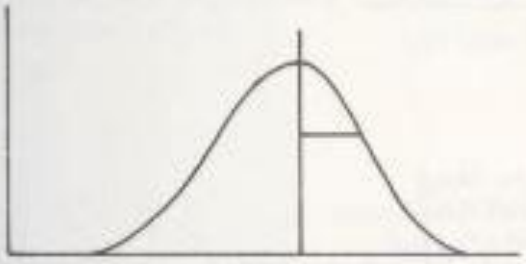
6.10 Substituting the results from Frame 16.2 directly in this equation we calculate

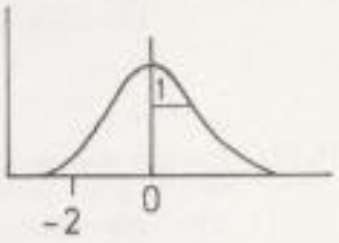
z = ?

$$\frac{96 - 100}{\frac{12}{\sqrt{36}}} = \frac{-4}{2}$$

= −2 again

6.11 Mark Mrs H's result in *the standard normal curve.*





16.12 What is the next step in performing any significance test?

Calculate p.

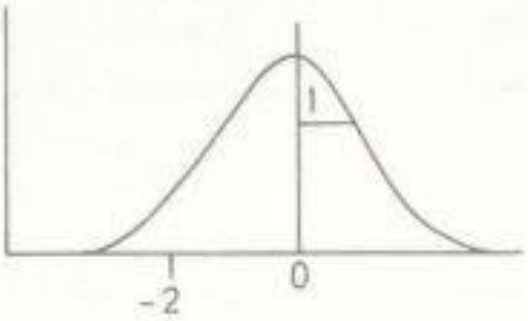
16.13 What does p represent?

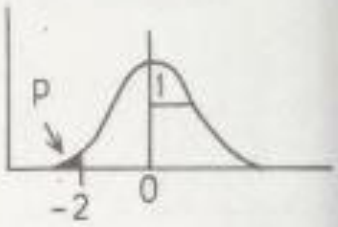
The probability of the result or a more extreme result arising by chance.

16.14 Look at Frame 16.2 again. Is Mrs H interested in both ends of the scale?

No. Only the end where her sample I.Q. is *lower* than normal.

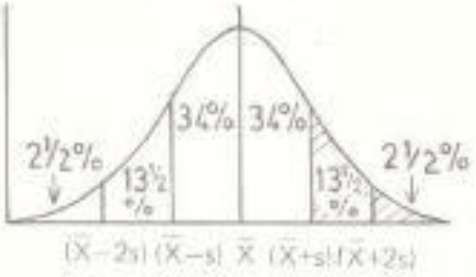
16.15 Shade the equivalent p area in this test (using the results in Frame 16.11)





16.16 What is the size of the area equivalent to this value p? The diagram in Frame 14.5 is repeated here.

.025 (2½)



16.17 Confirm this result from the tables relating p to z in the pull-out. The p-value there refers to one/two end(s) of the distribution.

Two.
p = .05 (past z = 2) at both ends together.
p = .025 (past z = −2) at one end.

16.18 p = .025 in this significance test. Is .05 > p > .01?

Yes.

6.19 What conclusion does Mrs H draw?

The result is significantly lower at the .05 significance level.
Children with bilharzia have a significantly lower I.Q. at this level.
IN.B. Significance tests say nothing about whether bilharzia *causes* the lower I.Q. only that there is in fact a relationship. Those with lower I.Q.'s may in fact have been more likely to contact the disease by swimming in infected water.)
There is as yet insufficient evidence for a significant difference at the .01 significance level.

6.20 You have completed your first significance test using 'z'. As z gets bigger the p value gets

.....

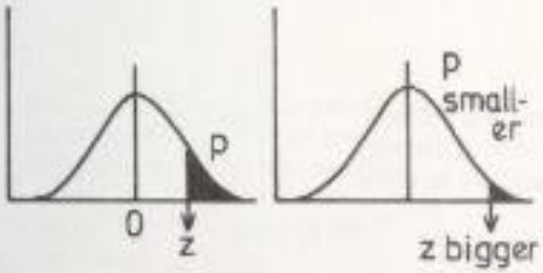
Smaller.

6.21 The Null Hypothesis is rejected if p is bigger/smaller than the significance level. Therefore the Null Hypothesis is rejected if the z you calculated from the results is bigger/smaller than the significant value of z

Smaller.

Bigger.

i.e.



6.22 If Mrs H had not known μ she would have estimated it using the mean of a ..... sample.

Control.

16.23 Of what would her control sample have consisted?

A random sample of simila children without bilharzia.

16.24 If Mrs H had not known σ she could have calculated ..... instead.

s.

16.25 State a formula she might have used

$$s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N - 1}} \quad \text{or}$$

$$\sqrt{\frac{\Sigma(X)^2 - \frac{(\Sigma X)^2}{N}}{N - 1}}$$

16.26 Look at these survey results. State the Null Hypothesis and its alternative.
Mrs H wants to know whether there is any difference in the mean weight of children aged 8 years with bilharzia compared with those without.
She calculated the mean weight of a random sample of 50 bilharzial children as 60.2 lb. and of a random sample of 50 non-bilharzial children as 62 lb. The standard deviation is 5 lb.

The Null Hypothesis states that any difference is due to chance whereas its alternative is that children with bilharzia have a different weight to normal children.

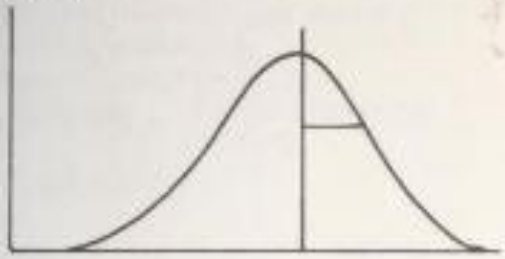
16.27 Assuming the Null Hypothesis is true, the sample of children with bilharzia and her control are presumed to come from the same .....

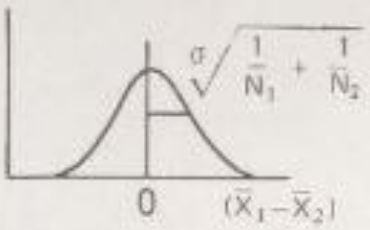
Population.

16.28 Therefore the difference between the bilharzial sample mean and the control mean follows which distribution?

The distribution of the difference between two sample means.

16.29 Complete this distribution in general terms





16.30 What is the value of

$$s\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}$$

using the data in Frame 16.26.

$$5\sqrt{\frac{1}{50}+\frac{1}{50}}=5\sqrt{\frac{1}{25}}=1$$

16.31 What is the value of $(\bar{X}_1-\bar{X}_2)$ in that Frame?
(Calling children with bilharzia the 1st sample and the control group the 2nd)

−1.8 lb.

16.32 This value is just 1 of many possible values of $(\bar{X}_1-\bar{X}_2)$ under the Null Hypothesis. How many standard deviations is it from the mean?

−1.8
That is its z value.

16.33 What is the formula for calculating z using the distribution of the difference between two sample means?
(Refer back to Frame 14.55 if you need.)

$$z=\frac{\bar{X}_1-\bar{X}_2}{\sigma\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}}$$

or

$$z=\frac{\bar{X}_1-\bar{X}_2}{s\sqrt{\frac{1}{N_1}+\frac{1}{N_2}}}$$

16.34 What is z calculated to be from the data in Frame 16.26.

$$z = \frac{-1.8}{5\sqrt{\frac{1}{50} + \frac{1}{50}}} = -1$$

i.e. the same result.

16.35 Mrs H in Frame 16.26 is concerned with one/two tails, so the significance level applies to one/two tails?

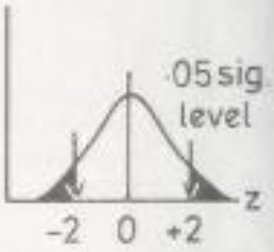
Two.
Two (you can use the table in the pull-out directly).

16.36 What is the equivalent significant value of z using the tables in the pull-out relating z to p

(a) if the significance level is .05

(a) For .05, the significant value of z = 2.0

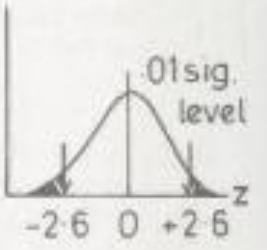
i.e.



(b) if the significance level is .01

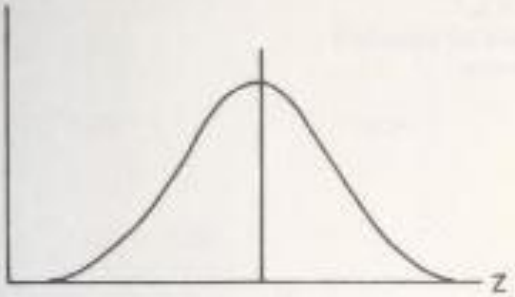
(b) For .01, the significant value of z = 2.6

i.e.

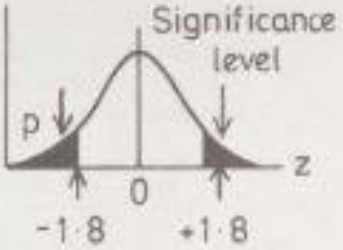


6.37 We decided that we would reject the Null Hypothesis if the z calculated from the results is bigger than the significant value of z.
Here calculated $z = -1.8$
Significant $z = 2.0$ (.05) or 2.6 (.01)

Sketch this idea.



What is your conclusion?



We accept the Null Hypothesis Either there is no difference or there is insufficient evidence.

6.38 Your calculated value of z from the experiment was bigger/smaller than the equivalent significant value z, so p was bigger/smaller than the significance level and so the Null Hypothesis was not rejected.

Smaller.

Bigger.

6.39 If you know the value of μ you do/do not need a control sample, and under the Null Hypothesis the sample mean follows which distribution?

Do not.

The distribution of sample means.

6.40 Under what condition?

If the sample is random and fairly large.

6.41 Then z can be calculated using which formula?

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}}$$

6.42 If you do not know μ you can estimate it from a control sample and then

$z = ?$

$$\frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

or

$$\frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

16.43 When you do not know σ use

.....

s from the samples.

16.44 There is a value of s^2 in each sample and it is usual to use both as follows. If the variance in the first sample is calculated to equal s_1^2 and the other variance is s_2^2 you can calculate the overall standard deviation using the formula.

$$s = \sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{?}}$$

$$\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}$$

N.B. when σ^2 is known and $s_1^2 = s_2^2 = \sigma^2$ this reverts to

$$\sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}$$

16.45 Look at this survey.
Professor T wishes to know whether there is any difference in width of a particular thoracic vertebra between different ethnic groups. He measures, using a standardised X-ray procedure, the relevant width on the random samples of 64 Zambians and 32 Portuguese East Africans. For the Zambians the mean width was 7.31 units (Variance .06) and for the Portuguese East Africans the mean was 7.16 units (Variance .05). Perform the necessary test by completing the schedule below.

State the Null Hypothesis

The difference is only due to chance.

State its alternative

There is a significant difference between the widths.

Under the Null Hypothesis these particular experimental results follow which distribution?

The difference between two sample means.

(contd. on opposite page)

6.45 *(contd.)*

You do not know σ^2 so you calculate s_1^2 and s_2^2 and use $z =$ (formula)

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

equivalent to

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

$\therefore$ z here = (value)

$$= \frac{7.31 - 7.16}{\sqrt{\frac{.06}{64} + \frac{.05}{32}}} = \frac{.15}{\sqrt{\frac{.16}{64}}} = 3$$

Professor T is interested in ..... tails(s)

2.

The equivalent z value to your significance level (pull-out) =

2.0 for a significant level of .05, 2.6 for a significant level of .01

Calculated z from the experiment is more/less than the significant value of z and p is more/less than the significance level.

More.
Less.

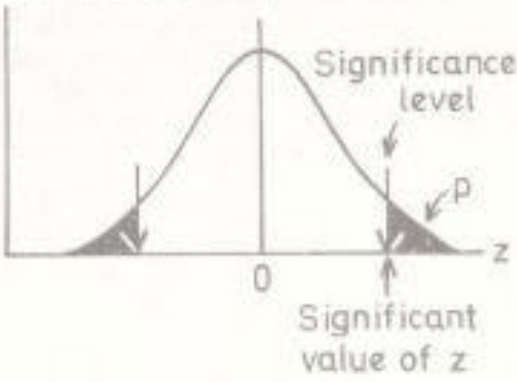
Therefore the conclusion is
.....
.....
.....
.....

That there is a significant difference between Zambians and Portuguese East Africans.
$(.01 > p)$.

6.46 **Practical Example:**

Repeat the format on the previous frame to decide whether the means here are significantly different.
Dr A wanted to know whether inclusion of B_{12} into guinea pig diets increased the weight gain. After a fixed period he found the mean weight gain of 50 randomly chosen guinea pigs without B_{12} was 5.2 oz. (s = 1.4) and of 50 randomly chosen guinea pigs given B_{12} was 5.5 oz. (s = 1.3).

16.47 You are possibly concerned about the tie-up between the *significance levels*, .05 and .01, and p on the one hand, and the *significant value of z* and *calculated z* on the other.
Does this diagram help?



Yes — I hope.
The significance levels .05 and .01 are equivalent to the significant value of z.
p and calculated z relate to the actual results calculated from the data.

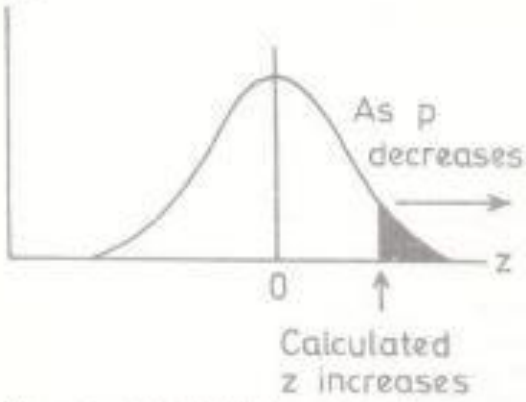
16.48 The experimental results themselves provide a value z called 'calculated z.' It is a measure of how extreme your particular result is, the Null Hypothesis being .....

True.

16.49 p is the probability of such an extreme or more extreme result of calculated z occurring by chance. The value of p increases as its calculated z. .....

Decreases.

16.50 i.e.,



You reject the Null Hypothesis if p is smaller/greater than the significance level, i.e., if calculated z is smaller/greater than the significant value of z.

Smaller,

Greater.

16.51 4 conditions for applying z tests should be satisfied before they can be used.

a) The sample must be chosen ..... Randomly.

b) The data must be qualitative/ quantitative. Quantitative.

c) The variable must be distributed in the population ..... Normally (although this is not important, in fact, if the samples are particularly large.)

d) The size of the sample must be 30 or greater (there is 1 exception we will mention again in the next chapter!)

16.52 We have used z tests on the data in Frames:

16.2
16.26
16.45
16.46

Were all these conditions satisfied? Yes.

16.53 In the next chapter we will solve problems where small samples are involved but the three conditions otherwise are as for z tests. State them.

Random samples.
Quantitative data.
Normal distribution.

SUMMARY

Simple tests involving 'z' are used.

1. When the samples are random.
2. When the data is quantitative.
3. Usually when the variable is normally distributed in the population.
4. Usually when the samples involved are bigger in size than 30.

(contd. on next page)

SUMMARY *(contd.)*

A. The first test is to test the difference between a sample mean and a know value of μ.

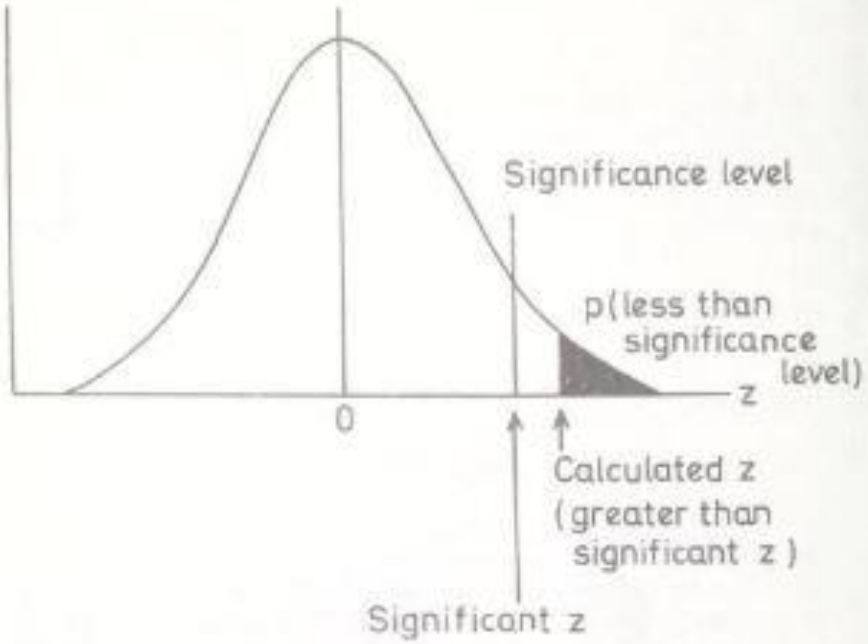
Here $z = \dfrac{\bar{X} - \mu}{\dfrac{s}{\sqrt{N}}}$ or $\dfrac{\bar{X} - \mu}{\dfrac{\sigma}{\sqrt{N}}}$ if σ is known.

B. The second test is to test the difference between the two sample means o a sample mean and a control mean.

Here $z = \dfrac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\dfrac{s_1^2}{N_1} + \dfrac{s_2^2}{N_2}}}$ if σ^2 is not known.

or $\dfrac{(\bar{X}_1 - \bar{X}_2)}{\sigma\sqrt{\dfrac{1}{N_1} + \dfrac{1}{N_2}}}$ if σ^2 is known.

As the calculated value of z increases, p, the probability of a more extreme value by chance decreases.



The Null Hypothesis is rejected if the significance level is *greater* than p. This is equivalent to the particular significant value of z being *smaller* than z calculated from the results.

Chapter 17

SIMPLE TESTS WITH STUDENT'S 't'

INTRODUCTION

The method for significance testing when the samples are smaller than 30, say, was discovered by a man called Gosset in 1908. At the time he was employed by Guinness Brewery in Dublin. The firm's regulations required him to use a pen-name and he chose the name 'Student'; 't' was the symbol later introduced in connection with the distribution used, which is consequently known as Student's 't'.

17.1 What are the criteria for using z?

Random samples.
Quantitative data.
Normal distribution.
Sample size at least 30.

17.2 Occasionally you can use the z tests even if N is less than 30. The requirement is that σ is known accurately and not estimated using .....

s

17.3 If N is less than 30 and σ is unknown, t tests are required.
Complete this table of tests to be used.

	N < 30	*N 30 or more*		
σ known	?	z	z	z
σ unknown	?	?	t	z

(In fact, t can be used whenever σ is unknown, but when N is bigger than 30, t becomes so like z that z can be used instead.)

17.4 The criteria for using t are otherwise the same as for z. State the criteria for using t tests.

Random samples.
Quantitative data.
Normal distribution.
Sample size less than 30 and σ unknown.

17.5 When t is used we never use ..... but s^2.

σ^2.

17.6 State the formulae for calculating s^2

$$\frac{\Sigma(\bar{X} - \bar{X})^2}{N - 1}$$

$$\frac{\Sigma(X^2) - \frac{(\Sigma X)^2}{N}}{N - 1}$$

17.7 Here is an example.
Would you use t or z?
Why?

Dr C is interested to know whether people who have had heart attacks have a blood cholesterol level different from the normal level of 180 mg/100 ml. He has a random sample of 16 patients and calculates their mean blood cholesterol as 195 mg/100 ml with a variance of 900.

t.
Random sample.
Quantitative data.
Blood cholesterol levels can be assumed to follow the the normal distribution in the population.
N is less than 30 and σ is unknown.

17.8 What are the stages in performing a significance test?

a) State the Null Hypothesis and its alternative.
b) Calculate p (z or t)
c) Draw conclusions.

17.9 Perform the first stage for the data in Frame 17.7.

The Null Hypothesis is that the difference is only due to chance.
The alternative is that people surviving heart attacks have a different blood cholesterol level.

17.10 A t distribution is very like the standard normal distribution. The formula for calculating t is very similar to that for calculating.....

z.

17.11 μ in Frame 17.7 is known.
Therefore if 30 or more patients had been used, which formula would you have used for z?

(Look back to the summary at the end of Chapter 16 if you have a memory like a sieve!)

$$z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}}$$

or $$z = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

as σ is not known in this example.

17.12 $$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

It doesn't equal $\dfrac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{N}}}$ Why?

Because if σ is known, t wouldn't be used.

17.13 What is the value of t in Frame 17.7?

$s^2 = 900 \therefore s = 30$

$$t = \frac{195 - 180}{\frac{30}{\sqrt{16}}} = \frac{15}{7\frac{1}{2}} = +2$$

17.14 Calculated t = 2. The next problem is to find the .....of t.

significant value.

17.15 This is not so straightforward as for z, as there are a series of t distributions which cannot all be standardised to one t distribution. They all depend on the symbol f.
f equals the value of the denominator when s^2 is calculated. What value has f in Frame 17.7?

f = 15 Because

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{N - 1}$$

$$= \frac{\Sigma(X - \bar{X})^2}{15} \text{ here.}$$

15 is the value of the denominator.

17.16 The significant value of t which you require is that where ..... = 15.

f = 15.

17.17 The t tables like the z tables include the area in one/two tails?

Two.

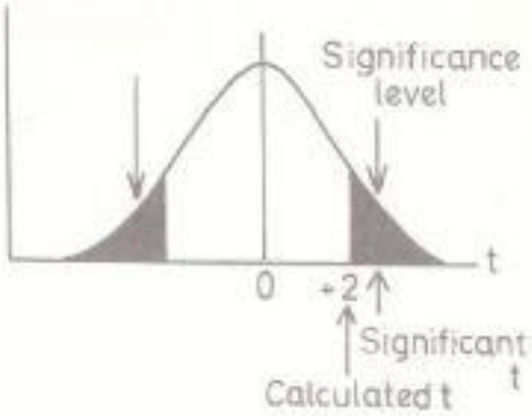
17.18 In Frame 17.7 Dr C is interested in one/two tails?

Two.

17.19 Therefore you need the significant value of t when f = 15 in the column of the significance levels for 2 tails.
Take a big breath and look at the t tables in the pull-out. What is the required significant value of t?

2.131 for .05
2.947 for .01

17.20 The conclusions for calculated and significant t are the same as they would have been had we a calculated z and a significant z.
What is your conclusion here?



In both cases calculated t is less than the significant t. You retain the Null Hypothesis . Your conclusion is either that there is no real difference or that there is a real difference and insufficient evidence.

17.21 Do you remember how you could distinguish between these 2 conclusions?

Yes, I hope.
Make N bigger.

17.22 Dr Clever Dick tells Dr C that in fact he should only have tested to see whether the blood cholesterol level was higher than normal. (The literature excludes the possibility of it being lower.) The test then would include one/two tails?

One only.

17.23 Remember that the t tables include both tails. What are the significant values of t now?

1.753 for .05
2.602 for .01

17.24 Dr Clever Dick's conclusion, if he had been performing this experiment, would have been what?
(Remember we calculated t to be 2)

He would have accepted the alternative as true at .05 but not at .01 ($.05 > p > .01$)
Actually he was quite clever to get a significant result.
This indicates how significance tests are made more sensitive if one tailed tests can be used.

17.25 When N = 20, f = ...

19.

17.26 As f gets bigger the t distribution becomes more nearly normal until when f = ..., we can safely use z tables instead of t.

f = 29
(N = 30)

17.27 f represents what is called the number of degrees of freedom (or free choices).
A playing captain wants to choose the rest of his hockey team.
In statistical symbols

..... = 11
(symbol)

N = 11

..... = 10
(symbol)

f = 10
Because the captain is playing, he only has 10 'degrees of freedom' (free choices).

17.28 This should give you some idea why statisticians pedantically call 'f', the number of degrees of freedom!

z does not use s^2 / σ^2 / f

z never uses f.

t does not use s^2 / σ^2 / f

t never uses σ^2.

17.29 What are the criteria for using t tests?

Random samples.
Data quantitative.
Normal distribution.
N less than 30 and σ not known.

17.30 Although t tests can be used to analyse results from very small samples these mini samples are not very sensitive. However, for the sake of practice, imagine here that the results refer to the pain threshold for only 5 random patients after a new analgesic. Are these results significantly higher than the population average, 4 units?
(I would not recommend such a small sample!)

Patient	A	B	C	D	E
Pain Threshold	7	5	2	4	7

$\bar{X}$ =

s^2 =

$\therefore$ s =

f =

N =

μ =

Calculated $t = \dfrac{\bar{X} - \mu}{\dfrac{s}{\sqrt{N}}}$

= ?

This is a ..... tailed test with f equal to ..... so the corresponding significant values of t =

What is your conclusion?

$\bar{X}$ = 5

$s^2 = \dfrac{\Sigma(X - \bar{X})^2}{N - 1}$

$= \dfrac{4 + 0 + 9 + 1 + 4}{4} = 4\frac{1}{2}$

$s \simeq 2.1$

f = 4 (The denominator when s^2 is calculated)

N = 5

μ = 4

$t = \dfrac{5 - 4}{\dfrac{2.1}{\sqrt{5}}} \simeq 1$

One

4

2.132 for .05
3.747 for .01

Calculated t is less than significant t. Still accept the Null Hypothesis. The analgesic is either ineffective or there is insufficient evidence.

17.31 Look at these results.

A psychiatrist got exasperated when patients 'phoned him during the night (waking him up!) to complain of insomnia. He heard of a new drug 'Zizz' and decided to try it randomly on 5 patients to see whether it was effective. He gave them alternately at random a placebo for a week and Zizz for a week and calculated the average number of hours each patient slept on the placebo and on Zizz.
Here are the results:

	Average on Placebo	*Average on Zizz*
Patient A	2	7
Patient B	6	13
Patient C	3	6
Patient D	1	0
Patient E	4	5

Incidentally Patient B lost his job!

They can be modified so that t can be used The pairs of results are subtracted so that X now refers to these differences. This modification is consequently called the paired-t-test (on differences).

Patient	*Average on Placebo*	*Average on Zizz*	*Difference (Zizz - Placebo)*
A	2	7	+5
B	6	13	+7
C	3	6	(a)
D	1	0	(b)
E	4	5	+1

a = +3

b = −1

17.32 These *differences* between the two drugs are based on a ..... sample of patients, the data is ....., the differences are ....., distributed, N. equals ....., and σ is/is not known.

Random
Quantitative
Normally
5 (the number of differences)
Is not.

17.33 Therefore we can use a ..... test on these *differences*. We treat these differences 5, 7, 3, −1, 1, as 5 values of a variable.

t

17.34 Perform the first stage.

The Null Hypothesis is that the differences are only due to chance. The alternati is that Zizz increases the average number of hours sleep.

17.35 We are going to perform the t test using these 5 differences (called X below).
Complete the table and calculate $\bar{X}$ and s for these differences.

Patient	X	$(X-\bar{X})$	$(X-\bar{X})^2$
A	5	2	4
B	7	4	16
C	3		
D	−1	−4	
E	+1		
	$\Sigma X =$	$\Sigma(X-\bar{X})=0$ of course	$\Sigma(X-\bar{X})^2 =$

Patient	X	$(X-\bar{X})$	$(X-\bar{X})^2$
A	5	2	4
B	7	4	16
C	3	0	0
D	−1	−4	16
E	+1	−2	4
	$\Sigma X = 15$		$\Sigma(X-\bar{X})^2 = 40$

$$\bar{X} = \frac{\Sigma X}{N} = ?$$

$$s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N-1}} = ?$$

$$\bar{X} = \frac{15}{5} = 3$$

$$s = \sqrt{\frac{\Sigma(X-\bar{X})^2}{N-1}} = \sqrt{\frac{40}{4}} = \sqrt{10}$$

17.36 $\therefore f = ?$

$f = N - 1 = 4$

17.37 Under the Null Hypothesis there is no real difference and so theoretically the mean difference should equal .....

0
i.e. $\mu = 0$

17.38 We use the same formula here as before.

$$\therefore t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

(Frame 17.12 if you'd forgotten).

17.39 In this example

$\bar{X}$ = ? (Frame 17.35)
μ = ? (Frame 17.37)
s = ? (Frame 17.35)
N = ? (the number of differences)

$\bar{X} = 3$
$\mu = 0$
$s = \sqrt{10}$
$N = 5$

$\therefore$ Calculated t =

$$\frac{3 - 0}{\frac{\sqrt{10}}{\sqrt{5}}} = \frac{3}{\sqrt{2}} \doteq 2.12$$

17.40 Calculated t = 2.12
The psychiatrist is interested in one/two tail(s).
The significant value of t = ...
(f = 4, remember).

One. He wants to know whether the drug is more effective than the placebo
2.132 for .05
3.747 for .01

17.41 Calculated t is ..... than the significant value of t. Therefore you accept/reject the Null Hypothesis.

Less.

Accept.

17.42 Such tests based on the difference between pairs of results on an individual are called paired-t-tests. In paired-t-tests the variable is X and we use

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

What do these symbols mean in the context of paired-t-tests?

X is the difference between the paired results.
$\bar{X}$ is the mean of the *differences*.
μ is the mean theoretical *difference*.
= 0
s is the standard deviation for these *differences*.
N is the number of *differences*.

17.43 Pairing results is a good idea if the results fall naturally into pairs i.e. each number is more closely related to its pair than any other result.
You have a series of twins. You should/should not arrange to treat the results in pairs.

Should.
A twin is more like its opposite number than any other person.

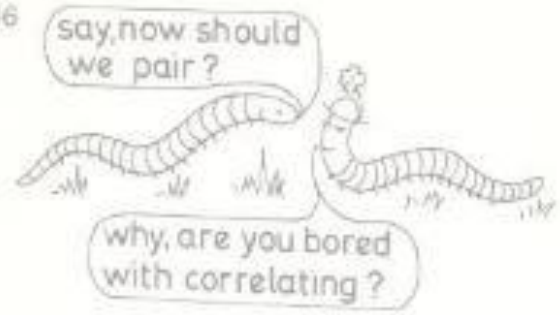
17.44 Noah, in his ark, had he had the time (or the inclination!) to do a drug trial on his animals, should/should not have used the paired-t-test?

Should.

17.45 When would you pair? (statistically!)

When each of the pairs is more like each other than the rest of the group.

17.46



17.47 Most frequently the paired-t-test is used when two drugs are given to each patient or when a specimen is tested using two different techniques. When results fall naturally into pairs we can treat the .....as the variable and straight away use

differences.

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}}$$

where μ = ?

0

17.48 Give two experimental situations where you would use the paired-t-test

a)

b)

For example:

a) Twins

b) A drug trial with 'before' and 'after' results.

c) In animal experiments where two animals are taken from each litter.

d) Some doctors match patients successfully for age/sex/severity of disease but if the matched people are not very alike this can be a mistake.

e) In comparing two chemical analytical methods.

17.49 To recap from the last chapter:

When z is used and μ is unknown we standardise the distribution of

.....

.....

.....

and use the formula:

z =

The difference between two sample means (one of which is from the control group usually).

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

17.50 ∴ When t is used and μ is unknown instead of using

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

You can imagine we use t = ?

σ is replaced by s

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

17.51 We have to modify our usual formula for s^2 so as to include the results from both samples. For z we could use each separately as in

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

We calculate s^2 as follows when using t, pooling the squares of the deviations from the means:

$$s^2 = \frac{\overset{1st\ sample}{\Sigma(X - \bar{X}_1)^2} + \overset{2nd\ sample}{\Sigma(X - \bar{X}_2)^2}}{N_1 - 1 + N_2 - 1}$$

Which formula for s^2, which you are used to, is this pooled formula most like?

$$s^2 = \frac{\Sigma(X - \bar{X})^2}{N - 1}$$

17.52 Let us take an over simplified example.

Suppose the 1st sample is 1, 2, 3.

X	$X - \bar{X}$	$(X - \bar{X})^2$
1		
2		
3		
$\Sigma X =$	$\Sigma(X - \bar{X}) = 0$	$\Sigma(X - \bar{X})^2 =$

What is N_1 ? $N_1 = 3$

What is $\bar{X}_1$? $\bar{X}_1 = 2$

What is $\Sigma(X - \bar{X})^2$ in the 1st sample? $\Sigma(X - \bar{X})^2 = 2$

Suppose the 2nd sample is 0,2,2,4.

X	$X - \bar{X}$	$(X - \bar{X})^2$
0		
2		
2		
4		
$\Sigma X =$	$\Sigma(X - \bar{X}) = 0$	$\Sigma(X - \bar{X})^2 =$

(contd. on opposite page)

17.52 *contd.*

What is N_2 ?
What is $\bar{X}_2$?
What is $\Sigma(X - \bar{X})^2$ in the 2nd sample?

$N_2 = 4$
$\bar{X}_2 = 2$
$\Sigma(X - \bar{X})^2 = 8$

$$\therefore \quad s^2 = \frac{\overbrace{\Sigma(X - \bar{X}_1)^2}^{\text{in 1st sample}} + \overbrace{\Sigma(X - \bar{X}_2)^2}^{\text{in 2nd sample}}}{N_1 - 1 \quad + \quad N_2 - 1}$$

$$= \quad ?$$

$$\therefore \quad s = \qquad$$

$$s^2 = \frac{2 \quad + \quad 8}{2 \quad + \quad 3}$$

$$= 2$$

$$s = \sqrt{2}$$

17.53 In the last chapter when we didn't know σ we used

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

We keep the two sample variances separate. To use t we must pool the information from the two samples and calculate one value s^2, which =

$$s^2 = \frac{\Sigma(X - \bar{X}_1)^2 + \Sigma(X - \bar{X}_2)^2}{N_1 - 1 + N_2 - 1}$$

17.54 Having calculated s^2 we use the formula

$t =$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}}$$

To be pedantic, unless the variances in both samples are approximately equal we do something else. This is rarely so, however, and so we can forget about this problem here.

17.55 f always equals the denominator when s^2 is calculated. Therefore to use t when μ is not known and the sums of squares of the deviations from the means are pooled.
f =

$N_1 - 1 + N_2 - 1$

$= N_1 + N_2 - 2$

17.56 Here are some fictitious results. The suggestion is that alcohol slows the reaction time. A very small sample of students was taken, 5 of whom were given a reasonable amount of alcohol and 3 drank Schhh . . . (by you-know-who!). As soon as an electric bell rang each was to press a button. The time lapse was recorded electronically. These are the reaction times:

Schhh . . .	*Alcoholic*
10 units	10 units
12 units	12 units
14 units	14 units
	16 units
	18 units

Perform the first step in the significance test.

The Null Hypothesis is that any difference is due only to chance.
The alternative is that alcohol slows the reaction time.

17.57 Complete this table and calculate your t value.

1st sample			*2nd sample*		
X	$X - \bar{X}$	$(X - \bar{X})^2$	X	$X - \bar{X}$	$(X - \bar{X})^2$
10			10		
12			12		
14			14		
			16		
			18		
$\Sigma X =$	0	$\Sigma(X - \bar{X})^2 =$	$\Sigma X =$	0	$\Sigma(X - \bar{X})^2 =$

$\bar{X}_1 =$ $\bar{X}_2 =$

$N_1 =$ $N_2 =$

$s^2 =$

$s =$

$f =$

$\bar{X}_1 = 12$ $\bar{X}_2 = 14$

$N_1 = 3$ $N_2 = 5$

$$s^2 = \frac{(4+0+4)+(16+4+0+4+16}{6}$$

$= 8$

$s = \sqrt{8}$

$f = 6$ (the denominator)

(contd. on opposite page)

7.57 *contd.*

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} = \frac{12 - 14}{\sqrt{8}\sqrt{\frac{1}{3} + \frac{1}{5}}} \triangleq -1$$

Calculated t $\triangleq$ −1

7.58 The suggestion is concerned with ..... tail(s).

One.

The significant values of t are under the columns headed ..... and ..... on the row f =

.10 .02
f = 6

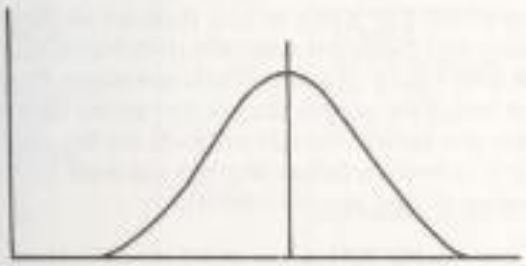
∴ Significant t from the t tables in the pull-out =

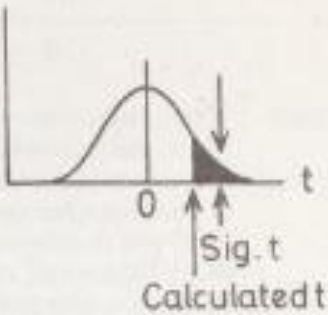
1.943 for .05
3.142 for .01

7.59 Significant t is bigger than calculated t.

Represent this fact in a sketch.

What is your conclusion?





There is no significant difference. Accept the Null Hypothesis. Either alcohol does not slow the reaction time or, if it does, there is insufficient evidence here.

7.60 State the conditions for using a t test.

Random samples.
Quantitative data.
Normal distribution.
N less than 30 and σ unknown.

17.61 The formula for calculating pooled s^2 where μ is unknown is

$$s^2 = \frac{\Sigma(X - \bar{X}_1)^2 + \Sigma(X - \bar{X}_2)^2}{N_1 + N_2 - 2}$$

Write this formula without using the means.

$$s^2 = \frac{\Sigma(X^2) - \frac{(\Sigma X)^2}{N_1} + \Sigma(X^2) - \frac{(\Sigma X)^2}{N_2}}{N_1 + N_2 - 2}$$

Use this result in the examp below.

17.62 **Practical Example**

In Frame 3.1 and 3.5 we have random samples of birth weights of children of diabetic and non diabetic mothers. Are the birth weights of children of diabetic mothers significantly bigger than the control group?
Most of the arithmetic you undertook in the Practical Example at the end of Chapter 7.

17.63 This schedule is included as a summary to help you decide the right test to use in the 3 following practical examples which are specially chosen so that you can practise deciding which z or t test to use. Assume all the results in the examples are random and based on normally distributed data.
First you decide whether σ or s is to be used. If σ is unknown the decision about which test to use is based on sample size, but in either case the next decision is whether μ is known (when the test depends on the distribution of the sample mean) or μ is unknown (when the test depends on the distribution of the difference of two sample means).

SUMMARY

Consider the questions in turn and decide on the test depending on the answers.

Is σ known?	Is N 30 or more?	Is μ known?	
If yes, use z tests		If yes use:–	$z = \dfrac{\bar{X} - \mu}{\dfrac{\sigma}{\sqrt{N}}}$
		If no, use:–	$z = \dfrac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\dfrac{1}{N_1} + \dfrac{1}{N_2}}}$
If no, z tests may be used in large samples	If yes, z tests are used:–	If yes use:–	$z = \dfrac{\bar{X} - \mu}{\dfrac{s}{\sqrt{N}}}$
		If no use:–	$z = \dfrac{\bar{X}_1 - \bar{X}_2}{\sqrt{\dfrac{s_1^2}{N_1} + \dfrac{s_2^2}{N_2}}}$
t tests in small	If no, t tests are used:–	If yes use:–	$t = \dfrac{\bar{X} - \mu}{\dfrac{s}{\sqrt{N}}}$ $(f = N - 1)$
		If pairs of differences ($\mu = 0$) use:–	$t = \dfrac{\bar{X} - 0}{\dfrac{s}{\sqrt{N}}}$ on the differences $(f = N - 1)$ N is the number of differences.
		If μ not known use:–	$t = \dfrac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\dfrac{1}{N_1} + \dfrac{1}{N_2}}}$ Where $s^2 = \dfrac{\Sigma(X - \bar{X}_1)^2 + \Sigma(X - \bar{X}_2)^2}{N_1 + N_2 - 2}$ and $f = N_1 + N_2 - 2$

(contd. overleaf)

SUMMARY *(contd.)*

1. Two students measured the length of the caecum in 25 male and 20 female specimens of a particular animal. They were interested to find out whether the caecal length was significantly different in the two sexes. They calculate the average male caecal length as 14.8 cm, and that for females 13.7 cm.
   What formula would they use to calculate s^2?
   What do the symbols in the formula represent?
   They calculated s^2 correctly to be equal to 0.81.
   What conclusion do they draw?
2. Professor X had the idea that people with cancer of the stomach ate more than others. He paired each of his 25 cases of cancer of the stomach with another patient with a different diagnosis but of the same age, sex, race and social class.
   He analysed the average daily intake and found that the mean difference was 180 calories. (Those with cancer eating more.)
   The standard deviation of the differences was 450 calories.
   What is his conclusion?
3. A Secretary for Health wanted to know whether a higher number of car accidents could be related to drivers with increased blood alcohol levels. He took blood samples of 100 random drivers involved in car accidents and the police chose 100 drivers randomly who had not been involved in an accident. For those involved in accidents the mean alcohol level was 2.42 units with a variance of 0.39. For the control group the mean was 2.24 units with a variance of 0.25.
   What conclusion would he draw?

Chapter 18

TESTING FOR REAL CORRELATION

INTRODUCTION

In Chapters 8 and 9 we learnt about correlation and how to calculate the correlation coefficients r and ρ. By chance a +ve or −ve value of the coefficient would usually be calculated even though in fact there existed no real correlation. Indeed it is exceedingly rare to obtain the exact result r or $\rho = 0$. This chapter shows you how to decide whether a particular value for r or ρ is likely to be due to significant correlation or chance variation from 0.

8.1 What does $1 - \frac{6\Sigma D^2}{N(N^2-1)}$ equal?

ρ

8.2 What is 'D' in the above equation?

The difference between rankings.

8.3 What is the name of the other correlation co-efficient you have met?

Pearson's Correlation Coefficient, r.

8.4 Suppose that you wished to decide whether a value of $r = +0.1$ represented real correlation or just a chance variation from $r = 0$.
What would your Null Hypothesis be?

That the variation from 0 was entirely due to chance.

8.5 If you were interested to detect real negative correlation this would involve a ..... tailed test.
Where no sign is specified a ..... tailed test is required.
The last frame required a ..... tailed test?

One.
Two.

Two.

8.6 To test your Null Hypothesis you use

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$$

(You needn't remember this formula).
What does N represent?

The number of pairs of results.

18.7 To use this formula r and ρ are interchangeable. What is the formula for testing whether a value of ρ represents real correlation?

$$t = \frac{\rho\sqrt{N-2}}{\sqrt{1-\rho^2}}$$

18.8 In this particular t test
$f = N - 2$
What does f represent?

f is the number of degress of freedom.

18.9 Suppose the correlation coefficient, +0.1 was calculated from 11 pairs of results —
$f = ?$

$N - 2 = 9$

18.10 $r = +0.1$
$N = 11$
$t = ?$

(Substitute in the formula in Frame 18.6)

$$t = \frac{+0.1\sqrt{9}}{\sqrt{1-.01}}$$

$$= \frac{+0.3}{\sqrt{.99}} \simeq +0.3$$

18.11 Calculated $t = +0.3$.
What does f equal again?
What are the corresponding significant t values in the 't' tables? (two tailed test).

f = 9.
t = 2.262 at .05.
t = 3.250 at .01.

18.12 Is calculated t bigger than significant t?

No. +0.3 is smaller than both 2.262 and 3.250.

18.13 Do you reject the Null Hypothesis?

No.

18.14 Do you conclude the correlation coefficient +0.1 here represents only chance variation from 0?

Yes, or that there is insufficient evidence.

18.15 Suppose that height and weight in your class are correlated with r = +0.6. You wished to test whether this represented real +ve correlation.

This is a .....tailed test.

One.

18.16 There were 27 members of the class measured so

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \quad = \quad ?$$

$$t = \frac{0.6\sqrt{25}}{\sqrt{1-0.36}} = \frac{3}{.8}$$

$$= 3.75$$

f = ?

f = 25

18.17 Calculated t = 3.75, significant t = ? (from the tables)

1.708 at .05
2.485 at .01 (1 tail)

18.18 What is your conclusion?

There is real +ve correlation between height and weight in the class.

18.19 Had only 6 members of the class been present – what would your conclusion have been? (Regarding the correlation coefficient, that is!).

$$t = \frac{0.6\sqrt{4}}{\sqrt{1-0.36}} = \frac{1.2}{.8} = 1.5$$

f = 4.

The conclusion would now be that the correlation was not real or that there was insufficient evidence.

18.20 **Practical Example**

What conclusion do you draw about your values for r and ρ calculated at the end of Chapter 9? Does the data represent real correlation?

SUMMARY

The formula used for testing the Null Hypothesis that there is no real correlation is

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \quad \text{or} \quad \frac{\rho\sqrt{N-2}}{\sqrt{1-\rho^2}} \quad \text{where } f = N-2.$$

Chat about which tests to use

Some readers have said that although they can understand the frames individually they would be confused about which tests they should use if faced with research data to analyse from scratch. This chat is intended for people who feel that this is a problem of theirs.

In the past some medical students at this University have undertaken small holiday research projects. Here are 4 modified examples. See whether you would now be able to perform the necessary tests. Remember, from the statistical point of view, it is as important to realise your limitations as to kn which tests are within your capabilities. You are not yet in a position to advise on all their projects.

Project 1

3 students, Messrs Berney, Makanza and Trachtenberg, measured the x-ray width of the 1st thoracic vertebra of 3 different groups of people. There were 100 in each group. They wished to know whether the 3 groups differed. Are you able to perform the applicable calculation?

No except to compare the groups two at a time using

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_1}}}$$

3 groups are properly compared together using the 'Analysis of Variance' technique which you do not know. (Incidentally, there was no difference found).

Project 2

Mr Jelbert measured:

(X) the average rise in the height of the diaphragm relative to the ribs; and,

(Y) the increase in the area of the heart and pedicle.

He had 60 of each measurements on a group of 60 x-rays. He wanted to know whether (X) was associated with (Y). If he had given you his results could you have given him the answer?

Yes, I hope.

Describe how you would have answered his problem.

$$r = \frac{\Sigma(XY) - \frac{\Sigma X \Sigma Y}{N}}{\sqrt{\left[\Sigma X^2 - \frac{(\Sigma X)^2}{N}\right]\left[\Sigma Y^2 - \frac{(\Sigma Y)^2}{N}\right]}}$$

then $t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}}$

Two tails, $t = 58$
(An association was found).

Project 3

Mr Shepherd wanted to compare the number of Kalanga Tribesmen with a whorl on the finger print of their right thumb, with the number in the Nanjanga tribe.
Could you work this out given the data?

No, those with whorls were counted, it is qualitative data.
This is the subject of the next Chapter.
(He found no difference.)

Project 4

Two exchange students from other Universities who have since graduated, Dr Arthur (Glasgow) and Dr Terry (Birmingham), measured the heights of (1) 35 eleven-year old schoolboys with goitres and (2) 30 eleven-year old schoolboys without goitres at a local mission school. They wanted to know whether the boys without goitres were bigger than those with. What formula would you use?

$$z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

This is a one or two tailed test?

One. (They found that boys without goitres were not significantly bigger.)

CONCLUSION

Although after completing this programme you are still limited in your knowledge of actual tests which are in use, the basic ideas are always those you know.
You should always understand what statistical tests are about and what statistical conclusions mean.

Chapter 19 SIMPLE TESTS WITH χ^2

INTRODUCTION

Most tests involving qualitative data depend on χ^2.

19.1 z, t, and r are all calculated from qualitative/quantitative data

Quantitative

19.2 What is the distinction between qualitative and quantitative data?

Qualitative data is counted. Quantitative is measured.

19.3 χ is written CHI and pronounced KI, χ^2 tests are called ..... tests

Chi-squared.

19.4 χ^2 depends on f, like .....
(Symbol)

t.

19.5 What does f represent?

The number of degrees of freedom.

19.6 Qualitative data is usually counted into groups or categories. An example is blood groups. The Null Hypothesis says that any variation between the observed number in the groups and what you would expect is due to what?

Chance variation only.

19.7 If there is a significant difference the variation is ..... than is expected by chance and this suggests that some other factor is involved.

More.

19.8 As with z and t, if the calculated value of χ^2 is bigger than the significant value you accept/reject the Null Hypothesis.

If χ^2 is bigger, p is smaller and the Null Hypothesis is rejected.

19.9 Like z and t tables, χ^2 tables are used directly in one/two tailed tests?

Two.

19.10 Look at the X^2 tables in the pull out. Like the t tables, the column headings are s..... l..... and the rows correspond to different values. of .....(symbol)

Significant levels

f.

19.11 Compare the structure of X^2 tables with t tables. What is the important difference?

X^2 tables tabulate values for significance levels of .99 and .95 as well as for .10, .05, .02 and .01.

19.12 The .99 and .95 levels correspond to the state of affairs where the observed results differ from the theoretical results *less* even than you would expect by chance. In 99 or 95 cases out of 100 such or a more extreme result would occur by chance.
What does this infer?

The possibility of cheating.

19.13 In fact Mendel's pea observations based on genetic theory differed *less* than you would have expected by pure chance, $p > .95$. Do you think Mendel cheated?
Hint: He was an Abbot!

He didn't. In fact, it was subsequently found that the Abbot's gardener knew the results the Abbot wanted and tried to please him!

19.14 The main criteria for applying X^2 are:-

a) The samples are ..... chosen
b) The data is .....
c) Ideally the lowest expected frequency in any group is not less than 5.

Randomly.
Qualitative.
N.B. Assume all the samples are random in this chapter.

19.15 One of the commonest reasons for using X^2 is to see whether actual counts comply with those expected on theoretical grounds. *(Goodness of fit to a theory.)* This is so with the example below based on genetic theory. Are all the criteria for applying X^2 satisfied here?

Yes. The 3 criteria listed in the previous frame are satisfied. The lowest expected frequency is 25.

(contd. overleaf)

19.15 *contd.*

A geneticist was interested to see whether two plants had the genotype Aa. He crossed them to see how close the progeny were to the theoretical ratio –

½ Aa : ¼ AA : ¼ aa.

There were 100 progeny and these were his results: (a random sample)

Genotype	*Number Observed (O)*	*Number Expected on Theory (E)*
Aa	53	50
AA	23	25
aa	24	25
Total	100	100

19.16 In the last frame state the Null Hypothesis and its alternative.

Null Hypothesis: The differences are only due to chance.
Alternative: The differences are more than could be reasonably expected by chance.

19.17 $\chi^2 = \Sigma \dfrac{\left(\begin{matrix}\text{observed} \\ \text{number}\end{matrix} - \begin{matrix}\text{expected} \\ \text{number}\end{matrix}\right)^2}{\begin{matrix}\text{expected} \\ \text{number}\end{matrix}}$

say $\Sigma \dfrac{(O - E)^2}{E}$

There is a value of $\dfrac{(O - E)^2}{E}$ for each class. Σ is .....

and means .....

Capital sigma

Add together.

19.18 In Frame 19.15

Calculated $X^2 = \frac{(53 - 50)^2}{50}$ for genotype Aa

\+ ? for genotype AA

\+ ? for genotype aa

= which value?

$+ \frac{(23 - 25)^2}{25} + \frac{(24 - 25)^2}{25}$

$= \frac{9}{50} + \frac{4}{25} + \frac{1}{25} = \frac{19}{50}$

= 0.38

19.19 In Frame 19.15

Calculated $X^2 = 0.38$

To find the corresponding significant value of X^2 we need to know .....

f

19.20 Where X^2 is used, as in Frame 19.15, to decide whether the actual results 'fit' some theory (in this case genetic) $f = k - 1$ where k is the number of classes.

In Frame 19.15

f = ...

2

= 3 - 1 (There are 3 classes or genotypes)

19.21 The research worker using X^2 is nearly always interested in both tails, i.e. he is interested in differences between Observed and Expected results in either direction. This is/is not the case in Frame 19.15

Is.
For a one-tail test to be applicable the research worker must be aware of which classes he expects to contain fewer members and which he expects to contain more.

19.22 X^2 tables record both tails as they stand.

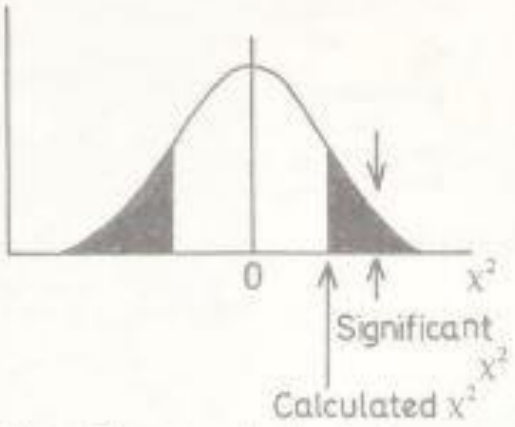
In Frame 19.15 f = 2

∴ Significant X^2 = ...

(two tails)

$X^2 = 5.991$ at .05
$X^2 = 9.210$ at .01
from the pull out.

19.23 Calculated χ^2 = 0.38 and is less than significant χ^2. What is your conclusion?



Hint: (The same as if calculated z or t was less than significant z or t).

You accept the Null Hypothesis. The conclusion is that the variation is insufficient to suspect that any other factor is involved and is due only to chance, i.e. the results 'fit' the genetic theory.

19.24 f = N − 1 in ordinary t tests where μ is known.
f = k − 1 in χ^2 testing "Goodness of fit" between actual results and results expected according to some theory.
N in these t tests is the number of ..... whereas in χ^2 k represents the number of .....

Results.
Classes.

19.25 f = N - 1 in the paired -t-test also.
Here N is the number of .....

Differences.

19.26 While we have been discussing this, the geneticist has performed another experiment. See below.
State the Null Hypothesis and alternative.

The Null Hypothesis is that it is only due to chance.
The alternative is that the variation is greater than that expected by chance.

(contd. on opposite page)

9.26 *contd.*

Here he expected the two factor segregation genetics ratio 9 AB: 3 Ab. 3 aB: 1 ab.

These are his results:

Phenotype	*Number Observed (O)*	*Number Expected (E)*
AB	245	
Ab	80	
aB	70	
ab	5	
Total	400	

9.27 There are 400 offspring in the last frame. The expected numbers in each class are in the ratio 9 : 3 : 3 : 1

= AB ?
Ab ?
aB ?
ab ?

225 AB 75 Ab 75 aB 25 ab

i.e. 9 : 3 : 3 : 1.

Use these expected values to obtain his conclusion in the next frame.

9.28 Can we apply χ^2 here?

Yes the lowest E = 25

The formula for χ^2 =

$$\Sigma \frac{(O-E)^2}{E}$$

Calculated χ^2 = +

$$\frac{(245-225)^2}{225} + \frac{(80-75)^2}{75}$$

+ +

$$+ \frac{(70-75)^2}{75} + \frac{(5-25)^2}{25} = 18\,^4/_9$$

= .....

f =

$k - 1 = 3$

(one/two tails)

Two tails

Significant χ^2 =

7.815 for .05
11.340 for .01

Conclusion =

Significant χ^2 is less than $18\,^4/_9$

Reject the Null Hypothesis. Conclude that there is more variation than you could reasonably expect by chance and that some further factor is involved (e.g. linking of genes).

19.29 So far the expected results were calculated on some theoretical grounds (Genetic). Just as sometimes for calculating z we use ..... to estimate σ^2, so sometimes for X^2 we use the observed results (O) to estimate (E).

s^2

We use the observed result like this in testing whether one factor is *associated* wi another.

19.30 When we use X^2 to test *association* rather than to test *goodness of fit* to a theory it affects the value of f. What does f represent?

The number of degrees of freedom.

19.31 The data to be tested for association is arranged in a 'contingency table'. Here is an example. Is this a table of 'O's or 'E's?

Observed results ('O's)

In a survey to help decide whether a particular inoculation had any protective properties the following results were obtained during an epidemic:

	Inoculated	*Not Inoculated*	*Row Totals*
Affected	5	55	60
Not Affected	95	145	240
Column Totals	100	200	300

19.32 State that Null Hypothesis and its alternative here.

The Null Hypothesis is that any association is only due to chance.
The alternative is that an association really exists between inoculation and incidence, inoculation protecting.

19.33 We assume initially that the Null Hypothesis/Alternative is true and on this basis calculate the expected results using the row and column total.

Null Hypothesis

9.34 In Frame 19.31 using the column totals we see that 100 out of the total 300 or 1/3 were inoculated. Assuming the Null Hypothesis is true and that inoculation is not really associated with the incidence of the disease we would expect 1/3 of those affected to have been inoculated/not inoculated.

Inoculated.
i.e. 1/3 of the people are inoculated and as this is assumed to have had no effect 1/3 of those affected would be inoculated.

9.35 But in Frame 19.31 we see that a total of 60 people are affected. Therefore we would expect that ..... of them had been inoculated.

1/3 or 20

9.36 Similarly 2/3 of the total are not inoculated and so you would expect 2/3 of those 60 affected, i.e. 40 people to be affected and inoculated/not inoculated.

not inoculated

9.37 As inoculation is assumed to have no effect and 1/3 are inoculated you would expect also 1/3 of the 240 not affected to be inoculated.

i.e. you would expect ..... inoculated, not affected people.

1/3 x 240 = 80

9.38 The expected results calculated in the last 3 frames are shown below. How many not inoculated not affected people would you expect?

	Inoculated	*Not Inoculated*
Affected	20	40
Not affected	80	?

2/3 of those not affected
i.e. 2/3 of 240 = 160

19.39 The contingency tables for the observed results and for the expected results are shown below:

Observed (O)

	Inoculated	*Not Inoculated*	*Row Total*
Affected	5	55	60
Not affected	95	145	240
Column total	100	200	300

Expected (E)

	Inoculated	*Not Inoculated*	*Row Total*
Affected	20	40	60
Not affected	80	160	240
Column Total	100	200	300

What do you notice about the row totals and column totals in each table?

They are the same in both tables.

19.40 Also notice that each expected result equals

$$\frac{\text{its row total} \times \text{its column total}}{\text{the overall total}}$$

e.g. for the inoculated affected group in Frame 19.39 the expected result

$$= \frac{? \times 100}{?} = 20$$

$$\frac{60 \times 100}{300} = 20$$

19.41 How can you calculate the expected frequencies in contingency tables?

Use the formula

$$\frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

19.42 Give the three criteria for applying X^2, *i.e.*

The data is ..... and the

samples are ..... and no expected

frequency is less than .....

qualitative.

random.

5.

19.43 Can χ^2 be applied to our inoculation data here?

'O'

	Inoculated	Not Inoculated	Row Total
Affected	5	55	60
Not affected	95	145	240
Column total	100	200	300

'E'

	Inoculated	Not Inoculated	Row Total
Affected	20	40	60
Not affected	80	160	240
Column total	100	200	300

Yes, no expected result is less than 5, if the samples are random.

19.44 Remember we were interested to see whether the inoculation protected against the disease. We expect, if this is so, to observe fewer inoculated affected people than expected. Are there?
Is this a one or a two tailed test?

Yes.
20 were expected but only 5 were observed in this group.

A one tailed test.

19.45 What is the formula for calculating χ^2?

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}$$

19.46 In Frame 19.43

$$\chi^2 = \frac{(5 - 20)^2}{20} + ?$$

$$+ ? + ?$$

$$= ?$$

$$+ \frac{(55 - 40)^2}{40} + \frac{(95 - 80)^2}{80}$$

$$+ \frac{(160 - 145)^2}{160}$$

$$= \frac{3375}{160} \doteq 21$$

19.47 When χ^2 was used to test 'Goodness of fit' to a theory f =

$k - 1$

19.48 In using X^2 to test association in a contingency table f = (r - 1) (c - 1) where r is the number of rows and c is the number of columns in the body of the table.

∴ in Frame 19.43, f = ?

(2 - 1) (2 - 1) = 1

19.49 i.e. There is 1 degree of freedom.

This is because if 1 expected result is calculated in a 2 rowed–2 columned contingency table, as the row and column totals are fixed, the rest of the numbers in the table cannot be chosen freely.

E.g. Complete this fictitious table

	B	*Not B*	*Row Total*
A	10	?	40
Not A	?	?	85
Column Total	50	75	125

	B	*Not B*	*Row Total*
A	10	40 - 10 = 30	40
Not A	50-10 = 40	75 - 30 = 45	85
Column Total	*50*	*75*	*125*

i.e. there is only 1 free choice (1 degree of freedom)

19.50 Anyway, to come back to the inoculation problem.
In frame 19.44 you decided this was a ..... tailed test.
In Frame 19.48 you calculated f to equal .....
∴ What is the required significant X^2 values in the table?

One.

One.

2.706 for .05

5.412 for .01

19.51 In Frame 19.46 X^2 was calculated from the contingency tables to equal 21. What is your conclusion?

The Null Hypothesis is rejected. Inoculation protects significantly (.01 > p)

19.52 When X^2 is used to test *'Goodness of fit'* to a theory (e.g. genetics) f = ?

k – 1 where k is the number of classes

19.53 For testing for *goodness of fit* to a theory, the theory itself is used to calculate the expected results. However, to use χ^2 to test for *association* we use the observed results to calculate the expected and f = ?

(r − 1) (c − 1) where r is the number of rows and c is the number of columns.

19.54 In testing for association we calculated the expected values using which formula?

$$\frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

19.55 Calculate the expected contingency table for these observed results.

'O'

	Blonde	*Brunette*	
Blue eyes	23	12	35
Green eyes	2	3	5
Brown eyes	15	45	60
	40	60	100

Using

$$E = \frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

i.e. 'E'

	Blonde	*Brunette*	
Blue eyes			
Green eyes			
Brown eyes			

the table appears:

14	21	35
2	3	5
24	36	60
40	60	100

What is the value of f for this contingency table? Can χ^2 be applied here?

f = (3 − 1) (2 − 1) = 2
No there are too few green eyed people.

19.56 We can adjust this by pooling the Blue and Green Eyed people.
i.e.

'O'

	Blonde	*Brunette*	
Blue or green eyes	25	15	40
Brown eyes	15	45	60
	40	60	100

What is the expected contingency table now?

'E'

	Blonde	*Brunette*	
Blue or green eyes			
Brown eyes			

16	24	40
24	36	60
40	60	100

What is f now?
Can χ^2 be applied here now?

f = (2 − 1) (2 − 1) = 1
Yes, the lowest E is now 16
(Pooling results sometimes enables χ^2 to be applicable)

19.57 $\chi^2 = \Sigma \frac{(O-E)^2}{E}$

= ? in the last frame?

$$\frac{(25-16)^2}{16} + \frac{(15-24)^2}{24} + \frac{(15-24)^2}{24} + \frac{(45-36)^2}{36}$$

$\triangleq 14$

19.58 Assuming this is a two-tailed test with f = 1
The significant value of χ^2 from the X^2 tables is ?

3.841 for .05
6.635 for .01

19.59 What is your conclusion?

Calculated X^2 is bigger. The Null Hypothesis is rejected. There is a significan association between eye colour and hair colour.

$(.01 > p)$

19.60 What is the value of f in a 3 row x 8 column contingency table for testing for association?

$(3-1)\ (8-1) = 14$

19.61 By completing the answers below decide whether you think that knowledge of bilharzia protects children from risking contracting the disease.
Here are the results obtained by Dr V.

	No Knowledge	Some Knowledge	Good Knowledge	Row Total
No Risk	20	40	20	80
Definite Risk	40	60	20	120
Column Total	60	100	40	200

State the Null Hypothesis and alternative, calculating the expected table using the formula:

$E =$

The Null Hypothesis is that any association is only due to chance.
The alternative is that there is an association between knowledge and risk.

$$E = \frac{\text{row total x column total}}{\text{grand total}}$$

(contd. on opposite page)

61 *contd.*

We find

	No Knowledge	Some Knowledge	Good Knowledge	Row Total
No Risk				80
Definite Risk				120
Column Total	60	100	40	200

24	40	16	80
36	60	24	120
60	100	40	200

We can/cannot use χ^2

Can. All 'E's are over 5.

χ^2 = which formula?

$$\chi^2 = \Sigma \frac{(O - E)^2}{E}$$

χ^2 = which value?

$$= \frac{(20-24)^2}{24} + \frac{(40-40)^2}{40} + \frac{(20-16)^2}{16} + \frac{(40-36)^2}{36} + \frac{(60-60)^2}{60} + \frac{(20-24)^2}{24}$$

$$= 2\,^7/_9$$

f = ?

$(2-1)\ (3-1) = 2$

This is a one/two tailed test?

One (Protection is specified)

Significant value of χ^2 =

4.605 for .05
7.824 for .01

The conclusion is

Either knowledge does not protect or there is insufficient evidence.

What should Dr V do?
He was thinking in terms of a Health Educator.

Increase the numbers observed to distinguish between these two possibilities or do not employ such a person.

62 If Dr V had calculated χ^2 as 0.019 (f = 2), what idea should have entered your head?

This value is less than the .99 and .95 significance levels. You should have suspected cheating.

19.63 What is the formula used for f if testing for association?

$f = (r-1)(c-1)$

19.64 What is the value of f if χ^2 is used for testing 'Goodness of fit' to a theory?

$f = k - 1$

19.65 What are the criteria for employing a χ^2 test?

The samples are random.
The data is qualitative.
No E is less than 5.

19.66 If some values of E are much less than 5 and the contingency table is large – how can you sometimes overcome this obstacle?

By pooling some classes as we did in Frame 19.56.

19.67 What is the formula for χ^2 ?

$$\chi^2 = \Sigma \frac{(O-E)^2}{E}$$

19.68 **Practical Example**

Mr L. P. wants to know whether malignancy is associated with the site of cerebral tumours.

His results were:

	Benign	*Malignant*	
Frontal lobe	21	19	40
Temporal lobe	28	2	30
Other lobes	51	29	80
	100	50	150

What conclusion would he draw?

SUMMARY

χ^2 - chi-squared - is the distribution used for testing data where:

1. The samples are random.
2. The data is qualitative.
3. There is ideally no expected value less than 5.

Calculated $\chi^2 = \Sigma \frac{(O-E)^2}{E}$ where O is observed and E an expected result.

(contd. on opposite page)

SUMMARY *(contd.)*

For two-tailed tests the .05 and .01 columns in the X^2 tables are used directly. For one-tailed tests columns .10 and .02 are used for significance levels of .05 and .01. If it is found that the calculated value of X^2 is less than the .95 and .99 values it suggests the possibility of cheating.
$f = k - 1$ where k is the number of classes if testing 'Goodness of fit to a theory'.

Where X^2 is used to test for associations in a contingency table, the expected results are calculated using

$$E = \frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

Here $f = (r-1)(c-1)$ – where r is the number of rows and c is the number of columns in the body of the contingency table.

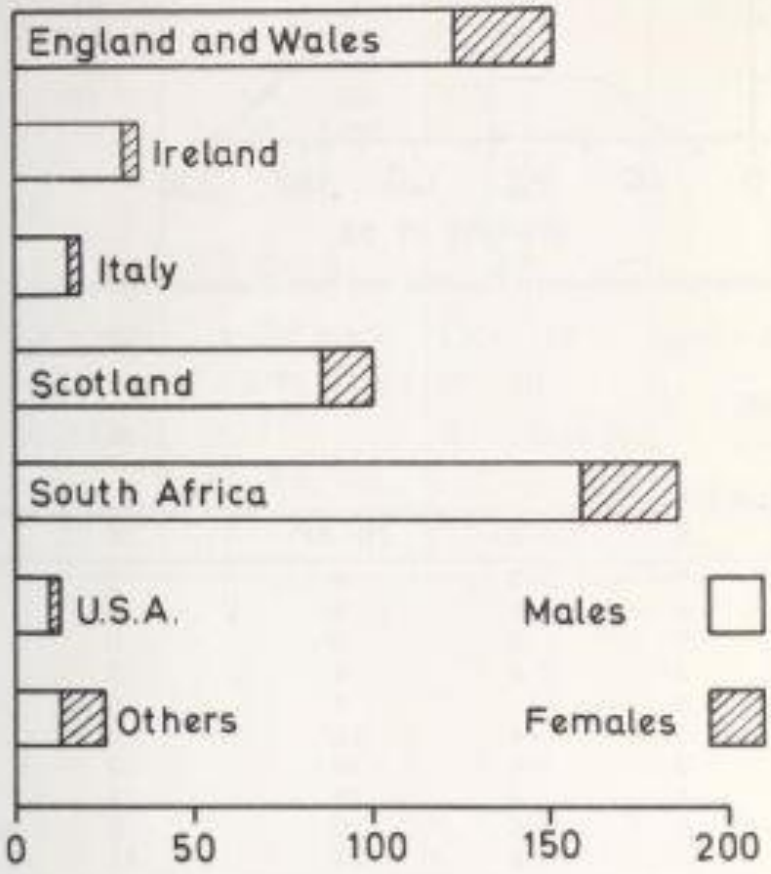
The Null Hypothesis states that the differences between the observed and expected results are only due to chance variation. If the calculated value of X^2 is greater than the significant value, the Null Hypothesis is rejected.

Note

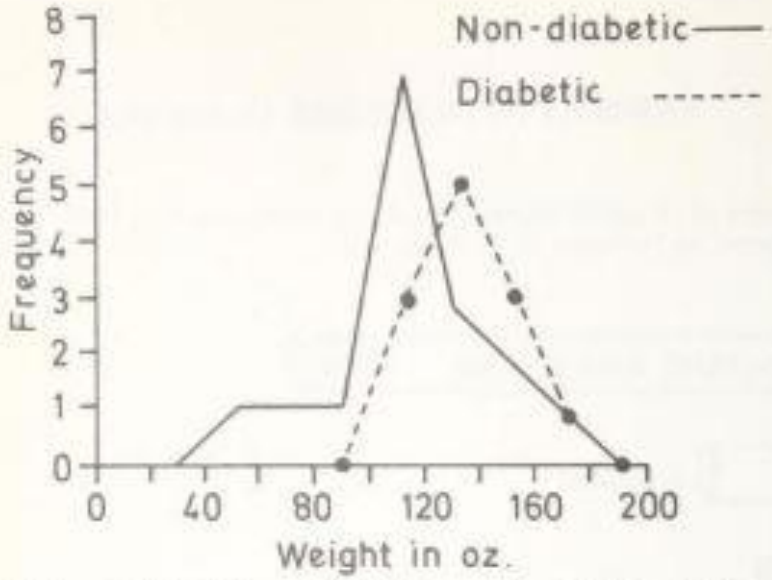
You have done well to complete this programme particularly if you have not sneaked a look at the answers before attempting to solve the frames yourself. This is the book concluded. Primarily you should be able to understand what most statistical jargon is about. I hope you can also perform simple tests for yourself. One of the aims is to have shown you what you cannot yet cope with. The keener people ought to be able to understand other statistical books by now. However, one of the best lessons to remember when doing research is that if in doubt as to how to analyse your data, and statistics is involved, ask advice *before* collecting the data.

ANSWERS TO PRACTICAL EXAMPLES

23 Country of Origin of Doctors Practising in Country X in 1967. Data Collected by Professor W. F. Ross, 1967.



3.29



Birthweight of children of Diabetic and Non Diabetic Mothers

5.29 For example:–

	X	$(X-\bar{X})$	$(X-\bar{X})^2$	X	(X^2)
1	2	−3	9	2	4
2	4	−1	1	4	16
3	5	0	0	5	25
4	3	−2	4	3	9
5	6	+1	1	6	36
6	9	+4	16	9	81
7	9	+4	16	9	81
8	1	−4	16	1	1
9	0	−5	25	0	0
10	11	+6	36	11	121
	$\Sigma(X) = 50$ $\bar{X} = 5$	$\Sigma(X) = 0$	$\Sigma(X-\bar{X})^2 = 124$	$\Sigma(X) = 50$	$\Sigma(X^2) = 374$

$$\therefore \frac{\Sigma(X-\bar{X})^2}{N-1} = \frac{124}{9} = 13\frac{7}{9}$$

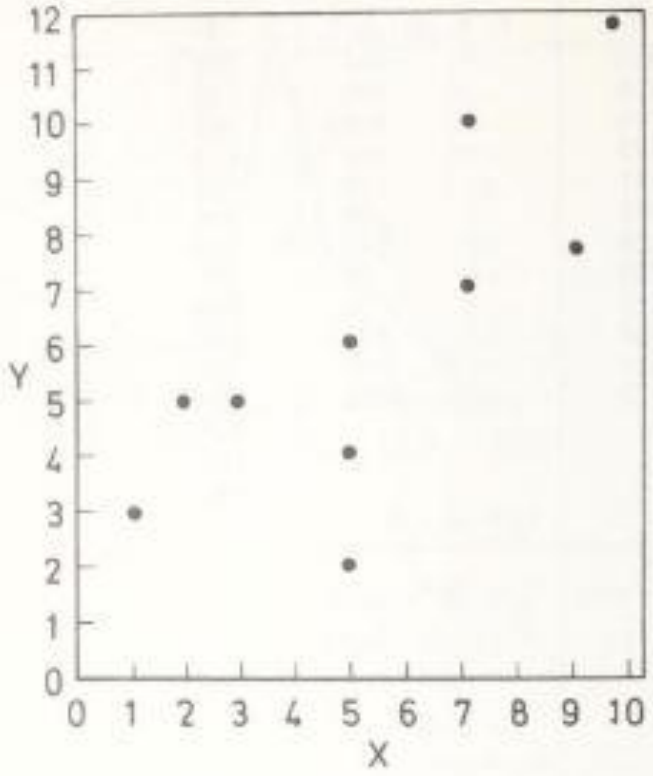
$$\therefore \frac{\Sigma(X)^2 - \frac{(\Sigma X)^2}{N}}{N-1} = \frac{374 - \frac{50^2}{10}}{9}$$

$$= \frac{374-250}{9} = \frac{124}{9} = 13\frac{7}{9}$$

42

Frame 3.1			Frame 3.3	
X	$X-\bar{X}$	$(X-\bar{X})^2$	X	X^2
103	−31	961	52	2704
114	−20	400	79	6241
114	−20	400	80	6400
122	−12	144	100	10000
131	−3	9	103	10609
138	+4	16	104	10816
138	+4	16	104	10816
138	+4	16	106	11236
143	+9	81	109	11881
146	+12	144	111	12321
151	+17	289	120	14400
170	+36	1296	121	14641
			127	16129
			149	22201
			150	22500
	$\Sigma(X-\bar{X}) = 0$		162	26244
$\Sigma X = 1608$	$\Sigma(X-\bar{X})^2 = 3772$		$\Sigma X = 1777$	$\Sigma X^2 = 209139$
$N = 12$	$s^2 = \frac{3772}{11} = 342.9$		$N = 16$	$s^2 = \frac{209139 - \frac{(1777)^2}{16}}{15}$
$\bar{X} = 134$	$\therefore s = 18.5$		$\bar{X} = 111.0625$	$= 785.4$
				$\therefore s = 28.0$

9.44 (1) Scatter Diagram



(2) Estimate of the Correlation Coefficient could be + 0.8

(3) Calculation of r.

X	Y	X^2	Y^2	XY
1	3	1	9	3
2	5	4	25	10
3	5	9	25	15
5	2	25	4	10
5	4	25	16	20
5	6	25	36	30
7	7	49	49	49
7	10	49	100	70
9	8	81	64	72
10	12	100	144	120
$\Sigma(X) = 54$	$\Sigma(Y) = 62$	$\Sigma(X^2) = 368$	$\Sigma(Y^2) = 472$	$\Sigma(XY) = 399$

4 *(contd.)*

$$N = 10$$

$$r = \frac{\Sigma(XY) - \frac{(\Sigma X)(\Sigma Y)}{N}}{\sqrt{\left(\Sigma(X^2) - \frac{(\Sigma X)^2}{N}\right)\left(\Sigma(Y^2) - \frac{(\Sigma Y)^2}{N}\right)}}$$

$$\therefore r = \frac{339 - \frac{54 \times 62}{10}}{\sqrt{\left(368 - \frac{(54)^2}{10}\right)\left(472 - \frac{(62)^2}{10}\right)}}$$

$$\therefore r = \frac{399 - 334.8}{\sqrt{(368 - 291.6)(472 - 384.4)}}$$

$$\therefore r = \frac{64.2}{\sqrt{(76.4)(87.6)}} = \frac{64.2}{\sqrt{6692.64}} = \frac{64.2}{81.8}$$

$$\therefore r = +0.78$$

(4) Calculation of ρ

X	Y	Ranked X	Ranked Y	D	D^2
1	3	10	9	+1	1
2	5	9	6½	+2½	6¼
3	5	8	6½	+1½	2¼
5	2	6	10	−4	16
5	4	6	8	−2	4
5	6	6	5	+1	1
7	7	3½	4	− ½	¼
7	10	3½	2	+1½	2¼
9	8	2	3	−1	1
10	12	1	1	0	0
		Sum to 55	Sum to 55	$\Sigma(D) = 0$	$\Sigma(D^2) = 34$

9.44 *(contd.)*

$$N = 10$$

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

$$\therefore \rho = 1 - \frac{6 \times 34}{10 \times 99}$$

$$\therefore \rho = 1 - \frac{34}{165}$$

$$\therefore \rho = 0.79$$

$$\therefore r = +0.78 \text{ and } \rho = 0.79$$

which are approximately the same.
Of course r is the more accurate estimate as it uses all the information.

11.62 You ought to have thought of most of the following points.

1) **Definition of the population**
The doctor must fully and exhaustively define his population so that tho to be included and excluded are obvious. For example is he only running his trials on adults, or females? How is he going to define overweight? Is I going to attempt to exclude people with renal or hormonal disease and if so, how?

2) **Factors affecting precision**
Is he going to weigh patients dressed only in a gown provided at the time What decisions is he going to make about diet? Over what period of time will he measure the decrease in weight? How many patients will he incluc

3) **Factors affecting bias**
The trial will fortunately be prospective and objective. Random samples must be allotted. If the patients are allotted numbers consecutively as they enter the trial the numbers can previously have been allotted to the different treatments using tables of random numbers. One group will be the control group on a placebo.

4) **Other Factors**
Check that the results will be analysable before starting (you will learn one test which is suitable soon). Decide before starting what will be done with patients who drop out of the trials. Decide what records of drug side-effects must be kept and what will be done with them.

63 This is a one-tailed test as the surgeon's wife was only interested in the physician's prowess at tossing tails.
The answer depends on the significance level.

For, the probability of tossing 1 tail is $\frac{1}{2}$ = .5

For, the probability of tossing 2 tails is $(\frac{1}{2})^2$ = .25

For, the probability of tossing 3 tails is $(\frac{1}{2})^3$ = .125

For, the probability of tossing 4 tails is $(\frac{1}{2})^4$ = .0625

For, the probability of tossing 5 tails is $(\frac{1}{2})^5$ = .03125

∴ Considering the .05 significance level p is less than .05 at 5 throws and she should have persuaded her husband to stop then.
However, for the significance level of .01, p only becomes less than this value at the seventh throw.

46 The Null Hypothesis states that there is no significant difference. The alternative is that B_{12} increases weight gain.

$$\text{Calculated } z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

$$\therefore \quad z = \frac{5.5 - 5.2}{\sqrt{\frac{1.3^2}{50} + \frac{1.4^2}{50}}}$$

$$\therefore \quad z = \frac{0.3}{\sqrt{\frac{3.65}{50}}} \simeq 1.15$$

This is a 1 tailed test.
The equivalent z values are 1.6 (0.5) and 2.3 (0.1).
Calculated z is less than these significant values of z and p is greater than the significance levels.

$(p > .05)$

Therefore either B_{12} does not increase the weight gain or there is insufficient evidence.

17.62 The Null Hypothesis states that there is no significance difference. It is a one-tailed test.

$\bar{X}_1$ for Diabetic Mothers is 134

$N_1 = 12$

$\bar{X}_2$ for Non-Diabetic Mothers is 111.0625

$N_2 = 16$

$\therefore \quad f = N_1 + N_2 - 2 = 26$

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{N_1} \quad \frac{1}{N_2}}}$$

$$\text{where } s^2 = \frac{\Sigma(X_1^2) - \frac{\Sigma(X_1)^2}{N_1} + \Sigma(X_2^2) - \frac{(\Sigma X_2)^2}{N_2}}{N_1 + N_2 - 2}$$

$$= \frac{(3772) + (209139 - 197358)}{26}$$

$$= \frac{15553}{26} = 598.2$$

$\therefore s = 24.5$

Note $\Sigma(X - \bar{X}_1)^2 = \Sigma(X^2) - \frac{(\Sigma X)^2}{N_1}$

$$\therefore t = \frac{134 - 111.0625}{24.5\sqrt{\frac{1}{12} + \frac{1}{16}}}$$

$= 2.45$

$\therefore$ Calculated $t = 2.45$

Tabulated t with $f = 26$ and using one tail is 1.706 for a significance level of .05 and 2.479 for the .01 significance level.

$\therefore .05 > p > .01$

$\therefore$ The conclusion is that the birth weights of children of diabetic mothers are significantly bigger than the control group at the .05 significance level but not at the .01

7.63 **Question 1**

σ unknown, $N < 30$, μ unknown

$$\therefore \text{Use } t = \frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \quad \text{where } s^2 = \frac{\Sigma(X - \bar{X}_1)^2 + \Sigma(X - \bar{X}_2)^2}{N_1 + N_2 - 2}$$

$\bar{X}_1 = 14.8$; $\bar{X}_2 = 13.7$; $s^2 = .81$; $N_1 = 25$; $N_2 = 20$

$$\therefore \text{Calculated } t = \frac{14.8 - 13.7}{.9\sqrt{\frac{1}{25} + \frac{1}{20}}} = \frac{1.1}{.27} \simeq 4$$

$f = N_1 + N_2 - 2 = 43$ which is not shown in the t tables in the pull-out. (If f = 29 or bigger the value of t does not change much and we use the bottom line).

This is a two-tailed test.

Significant t = 2.000 (.05) or 2.600 (.01)

$\therefore$ Calculated t $>$ Significant t

$\therefore$ There is a significant difference between the results ($.01 > p$)

Question 2

σ unknown, $N < 30$, Paired results (matched)

$$\therefore \text{Use } t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{N}}} \quad \text{on the paired results.}$$

$\therefore \bar{X} = 180 \quad s = 450 \quad N = 25 \quad \mu = 0$

$$\therefore t = \frac{180 - 0}{\frac{450}{\sqrt{25}}} = 2$$

This is a one tailed test with f = 24

$\therefore$ Significant t = 1.711 (.05) and 2.492 (.01)

These results can be summarised $.05 > p > .01$.
People with cancer of the stomach ate significantly more at .05 but not at .01 significance level.

Question 3

σ unknown $N > 30$, μ unknown

$$\therefore z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}}$$

$\bar{X}_1 = 2.42$ $\bar{X}_2 = 2.24$ $s^2_1 = .39$ $s_2^2 = .25$

$N_1 = 100$ $N_2 = 100$

$$\therefore \text{Calculated } z = \frac{2.42 - 2.24}{\sqrt{\frac{.39}{100} + \frac{.25}{100}}} = \frac{.18}{.08} = 2.25$$

This is a one-tailed test

$\therefore$ Significant z = 1.6 (.05) or 2.3 (.01)

i.e. $.05 > p > .01$

At the .05 level you accept that the rate of accidents was significantly affected by the alcohol level but at the .01 level you conclude that either there is no effect or insufficient evidence.

18.20 The Null Hypothesis is that there is no real correlation. It is a two-tailed test.

Tabulated t ($f = 8$) is 2.306 (.05) or 3.355 (.01)

$$\text{Calculated } t = \frac{.78\sqrt{8}}{\sqrt{(1 - .78^2)}} = \frac{.78 \times 2.8}{\sqrt{(1 - .6084)}} = \frac{2.18}{.626}$$

$$= 3.5$$

Calculated t is bigger than significant t at both levels ($.01 > p$)
The conclusion is that correlation is real.

These results may well be summarised in a medical journal:—
"r = +0.78. This is evidence of real correlation ($t = 3.5$ $.01 > p$)"
(0.78 may be substituted for ρ to reach the same conclusion).

19.68 The Null Hypothesis is that any difference is due entirely to chance. It is a two-tailed test.

Using Expected result = $\frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$

we calculated the Contingency Table for Expected values as:-

	Benign	Malignant	
Frontal Lobe	$26\frac{2}{3}$	$13\frac{1}{3}$	40
Temporal Lobe	20	10	30
Other Lobes	$53\frac{1}{3}$	$26\frac{2}{3}$	80
	100	50	150

No Expected result is less than 5 so we can calculate $\chi^2 = \Sigma\frac{(O-E)^2}{E}$

O	E	O−E	$(O-E)^2$	$\frac{(O-E)^2}{E}$
21	$26\frac{2}{3}$	$-5\frac{2}{3}$	$32\frac{1}{9}$	1.2
28	20	+8	64	3.2
51	$53\frac{1}{3}$	$-2\frac{1}{3}$	$5\frac{4}{9}$	0.1
19	$13\frac{1}{3}$	$+5\frac{2}{3}$	$32\frac{1}{9}$	2.4
2	10	−8	64	6.4
29	$26\frac{2}{3}$	$+2\frac{1}{3}$	$5\frac{4}{9}$	0.2
				13.5

$\therefore \Sigma\frac{(O-E)^2}{E} = 13.5$

Tabulated χ^2 $(f = (r-1)(c-1) = 2)$ = 5.991 for a significant level of .05 and 9.210 for a significance level of .01. Calculated χ^2 is greater than Tabulated χ^2. The conclusion with both significance levels is that there is an association between malignancy and the site of cerebral tumours ($.01 > p$)

HOW MUCH HAVE YOU LEARNT?

This is to test your Knowledge – the answers are given at the end of the test.

Questions *Answers*

1. a) Give an example of quantitative (or continuous) data.
   b) Give an example of qualitative (or discrete) data.
   c) Give two reasons why this distinction is important in statistics.

2. Of 200 births in the Lady Chatterly's Maternity Home last year, 90 were female.
   a) What was the ratio of males to females?
   b) What was the proportion of females?
   c) What was the percentage of females?

3. Is

$$\frac{\text{the number of sailors killed at Trafalgar}}{\text{the number of sailors involved at Trafalgar}}$$

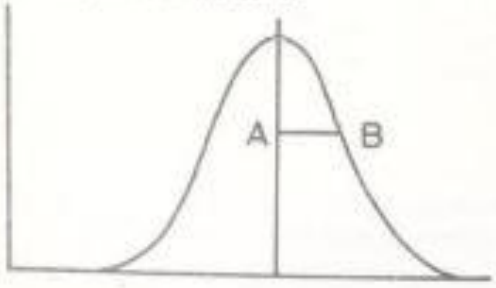
   a ratio, a rate, a proportion or a percentage?

4. Make a rough sketch of a) a histogram and b) a pie diagram

(a) Histogram (b) Pie diagram

5. When would you use a frequency polygon to present data?

6. This is a distribution of data.
   a) What is it called?
   b) Mark in the position of the mean.
   c) What is the length of AB called?
   (B is a point of inflection)



7 What is the difference between $\Sigma(X^2)$ and $(\Sigma X)^2$?

8 Calculate the mean, median and mode of the following distribution:-
1, 2, 2, 2, 3, 5, 5, 6, 6, 18.

9 Why is the mean a better measure of the middle than either the median or the mode?

0 Why is the sum of the deviations from the mean not used as a measure of variation?

1 Σ means "the sum of".

X is an observed result.

$\bar{X}$ is the mean.

s is the standard deviation.

N is the number of results.

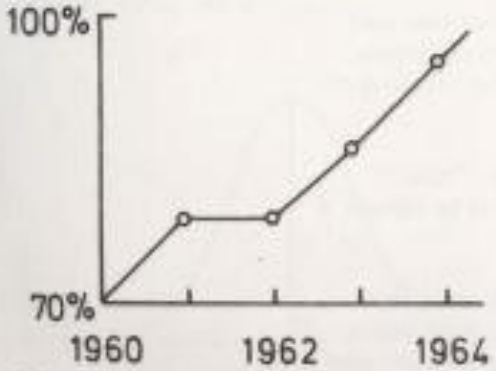
What is wrong with the following equation?

$$s = \frac{\Sigma(X - \bar{X})^2}{N - 1}$$

2 The variance of 1, 2, 3, 3, 4, 5 is 2.
What value has the standard deviation?
What value has the range?

3 Which is the better measure of variation, the range or the standard deviation? Why?

14 Why is this not a good diagram?

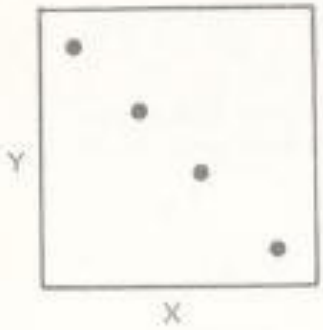


Percentage Pass rate in Anatomy at this Medical school (1960 - 64 inclusive)

15 What do you understand by "correlation"?

16 What are the maximum and minimum numerical values of a correlation coefficient? What is the value when there is no correlation?

17 The diagram below shows corresponding values for X and Y.
What is the value of the correlation coefficient here?



18 What is the main difference between Pearson's Correlation Coefficient and Spearman's?

19 ρ is Spearman's Correlation Coefficient

$$\rho = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$$

N is the number of pairs of results.
What does D represent?

20 s is a statistic and σ is a parameter and both represent the standard deviation.
What is the difference in the meaning of these 2 symbols?

21 What do you understand by "bias"?
What method would you use to reduce bias?

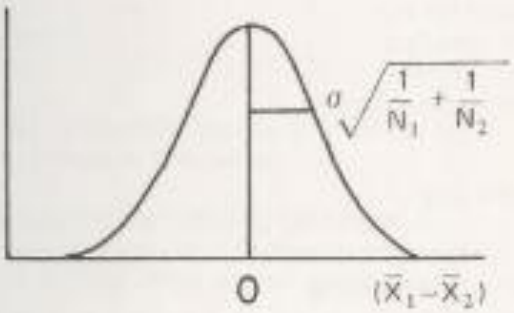
22 Precision can be used to describe how close various estimates of a population mean are to each other. How would you improve the precision of a sample?

23 You are interested in the I.Q's of this year's 1st Year Medical Students at your University as opposed to students in other faculties. Define your control group exactly.

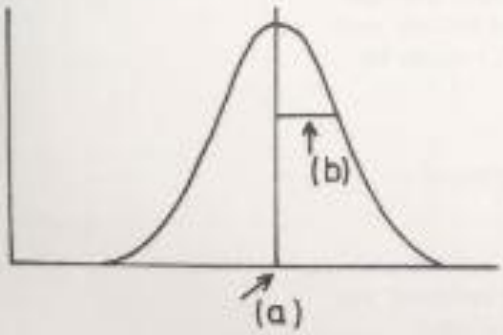
24 You are told that the probability equals one.
What does this mean?

25 a) What is the probability of throwing a 2 or a 6 with a "dice"?
b) What is the probability of throwing a 2 and a 6 with two dice?

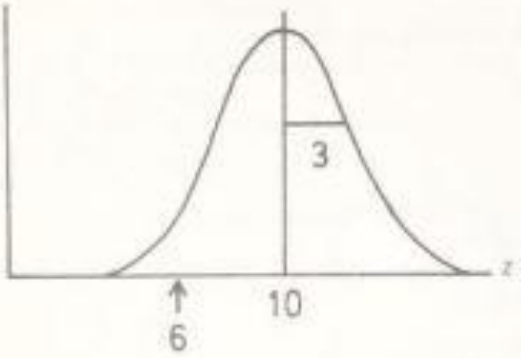
26 What type of sampling distribution is this?



27 There are many possible normal distributions. These can be standardised to a single normal distribution called the *standard* normal distribution (shown below). Label the arrows.



28 What is the value of z, the standard normal deviate, corresponding to the number, 6, in this diagram?



29 a) What is the special name for the standard deviation of the sampling distribution of the mean? b) How is it related to the population standard deviation?

30 What is the use of a significance (or confidence) level?

31 What do you understand by the meaning of the term "Null Hypothesis"?

32 An article in a journal states "$p > .05$". What conclusion would you draw about the Null Hypothesis?

33 What makes you decide to use a one as opposed to a two-tailed test?

34 Give the difference between the uses of z, the standard normal deviate, and Student's t (t - test) with respect to

a) the sample size

b) s^2 and σ^2

c) the number of degrees of freedom, f?

35 If you know the formula for calculating s^2 how can you use it to calculate the number of degrees of freedom?

36 Look at the pull-out.

What is the significant value of t at the 5% significance level (2 tailed test) where f = 3?

37 What conditions have to be satisfied for the application of X^2 (chi square)?

38 This is a 2 x 2 contingency table of observed results.

	A	Not A	
B	10	20	30
Not B	30	40	70
	40	60	100

a) Complete the equivalent expected table

	A	Not A	
B			
Not B			

b) How many degrees of freedom are there in this table?

39 Look at the pull-out. Calculated X^2 (chi square) with 11 degrees of freedom is 18.002. What is your conclusion?

a) in a 1 tailed test with the significance level .05.
b) in a 2 tailed test with the significance level .01.

40 What does a p value of .99 probably mean in a X^2 test?

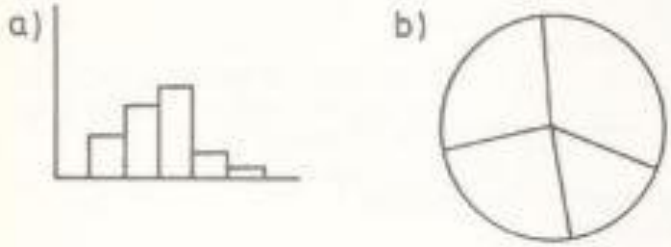
ANSWERS

1 a) Haemoglobin level. Bladder volume.
b) Sex. Blood groups.
c) The different types of data are presented differently and are subjected to different tests.

2 a) $\frac{110}{90}$ or $\frac{11}{9}$ 2 b) $\frac{90}{200}$ or $\frac{9}{20}$ 2 c) 45%

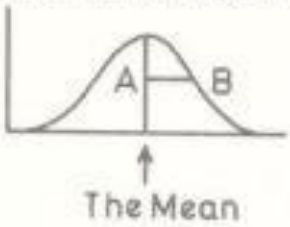
3 Proportion.

4



5 When the data is quantitative, especially when two sets of data are to be illustrated on the same diagram.

6 a) The normal distribution.
b)



c) The standard deviation.

7 $\Sigma(X^2)$ is the sum of the numbers already squared.
$(\Sigma X)^2$ is the square of the numbers already summed.

8 The mean = 5
The median = 4
The mode = 2

9 It uses all the information (and can be used further in significance tests.)

10 It always equals zero.

11 $s^2 = \frac{\Sigma(X - \bar{X})^2}{N - 1}$ or $s = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N - 1}}$

12 The standard deviation = $\sqrt{2}$
The range = 4

13 The standard deviation because it uses all the information (and can be used further in significance tests.)

14 The zero is suppressed and the line is extrapolated.

15 Association.

16 The maximum value = +1
The minimum value = −1
When there is no correlation the value = 0

17 −1

18 Pearson's uses the actual results and Spearman's the ranks.

19 D is the difference between ranks.

20 s is the standard deviation in a sample, σ is the standard deviation in the population.

21 Bias is the off-target effect of statistics.
Randomisation, "blind" sampling are examples.

22 Increase its size.

23 A random sample of 100's of this years 1st year students in other faculties at your university.

24 It is inevitable.

25 a) $\frac{1}{3}$ b) $\frac{2}{36}$ or $\frac{1}{18}$

26 The distribution of the difference between two sample means.

27 a) The mean = 0 b) The standard deviation = 1

28 $-1\frac{1}{3}$

29 The standard error.
It equals the population standard deviation divided by $\sqrt{N}$

30 It enables decisions to be made.

31 Any apparent difference is due only to chance variation.

32 It is either true or there is as yet insufficient evidence to reject it.

33 The alternative hypothesis is being only concerned with one outcome.

34 t is used

a) with small samples (say N less than 30)
b) with s^2 and
c) depends on the number of degrees of freedom, f.

z is used

a) with large samples (say N more than 30) and with smaller samples only if σ^2 is known.
b) It can be used with s^2 or σ^2 in large samples and
c) It does not depend on the number of degrees of freedom.

35 It equals the denominator.

36 3.182.

37 The samples are random.
The data is qualitative.
There is ideally no expected value less than 5.

38 a)

	A	Not A	
B	12	18	30
Not B	28	42	70
	40	60	100

b) f = 1

39 a) Reject the Null Hypothesis, Accept the Alternative.
b) Either the Null Hypothesis is true or there is insufficient evidence to reject it

40 Suspect cheating.

Printed by Photolithography by T. & A. Constable Ltd., Edinburgh